Reims Image 2014 - Tome 3 - CORESA

Yannick Rémion, Laurent Lucas

Tome 3 - CORESA

Reims Image 2014

Association Française d'Informatique Graphique (AFIG) / COmpression et REprésentation des Signaux Audiovisuels (CORESA) / Association Française de Réalité Virtuelle (AFRV) / Géométrie Discrète (GeoDis)

Reims, 25 novembre : journée des jeunes chercheurs IUT de Reims
26 à 28 novembre : centre des congrès de Reims

1res Journées plénières du GdR Informatique Géométrique et Graphique, Réalité Virtuelle et Visualisation

Reimsimage2014.univ-reims.fr

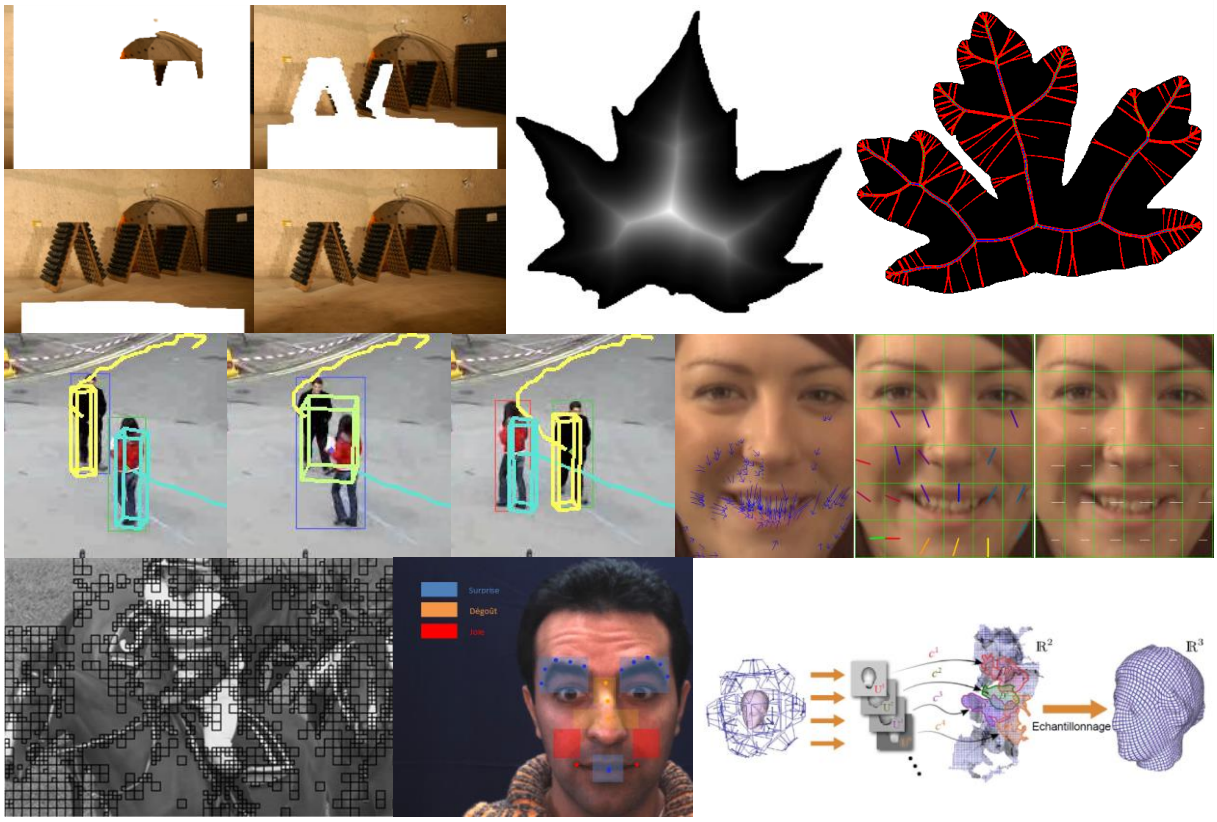
REIMS IMAGE 2014

UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

17E COLLOQUE CORESA

COMPRESSION ET REPRESENTATION DES SIGNAUX AUDIOVISUELS

26, 27 ET 28 NOVEMBRE 2014
CENTRE DES CONGRÈS, REIMS



Organisé par :

UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

CReSTIC

IUT REIMS CHALONS CHARLEVILLE

MAISON DE LA SIMULATION de Champagne-Ardenne

En partenariat avec

et le soutien de

Préface

Succédant aux différentes villes ayant accueilli les précédentes éditions (Lyon en 2010, Lille en 2012 et Le Creusot en 2013 pour les plus récentes), c'est Reims qui est l'hôte en novembre 2014 de la 17e édition du colloque CORESA (COmpression et REprésentation des Signaux Audiovisuels), organisé par le Centre de Recherche en Sciences et Technologies de l'Information et de la Communication (CReSTIC) et la Maison de la Simulation (MaSCA) de l'Université de Reims Champagne-Ardenne.

Ce nouveau millésime de CORESA offrira cette année encore aux chercheurs et aux praticiens dans le domaine du multimédia des sessions plénières scientifiques et techniques, des présentations orales, des posters, des démonstrations et des discussions sur des questions pertinentes et stimulantes concernant le futur du multimédia. Les thèmes abordés sont au cœur de la révolution multimédia : les nouvelles applications et nouveaux services qui naissent de la convergence entre les technologies et les usages des télécommunications, de l'audiovisuel et de l'informatique.

Lors de cette édition, nous avons reçu 20 soumissions : 13 articles ont été acceptés en sessions orales et 6 en session poster. Chaque article a été évalué par au moins 2 relecteurs. Cette édition 2014 est par ailleurs très spéciale puisque nous nous joindrons, pour la première fois, à d'autres communautés proches de la notre à savoir, l'AFIG (Informatique Graphique), l'AFRV (Réalité virtuelle, Réalité mixte) et GéoDIS (Géometrie Discrète).

CORESA 2014 est par ailleurs heureuse d'accueillir quatre chercheurs reconnus au niveau international dans des domaines d'activité en lien avec le multimédia :

- Mathieu Desbrun (California Institute of Technology - Caltech), « The power of primal/dual meshes »
- Edmond Boyer (INRIA Grenoble Rhône-Alpes - MORPHEO), « Modélisation des formes en mouvement »
- Natalya Tatarchuk (Engineering Architect - Bungie), « Applied graphics research for video games: solving real-world problems under real-world constraints »
- Frédéric Dufaux (Télécom ParisTech - LTCI), « Vidéo 3D – Technologies existantes et émergentes »

Enfin, afin d'encourager les jeunes doctorants un prix est attribué à la meilleure communication.

Nous espérons que le travail réalisé par le comité vous donnera entière satisfaction.

Bon séjour à Reims et Champagne !

Yannick Remion et **Laurent Lucas**
Université de Reims Champagne-Ardenne
Responsables du colloque CORESA 2014 – Reims Image 2014

Comité de pilotage

Président
• William Puech (LIRMM, Université Montpellier 2)

Membres
• Atilla Baskurt (LIRIS, INSA Lyon)
• Vincent Charvillat (IRIT, ENSEEIHT)
• Mohamed Daoudi (TELECOM Lille1 / LIFL)
• Gwenaël Doërr (Technicolor R&D France)
• Florent Dupont (LIRIS, Université Claude Bernard Lyon 1)
• Christine Guillemot (INRIA)
• Stéphane Pateux (Orange Labs)
• Carole Thiebaut (CNES)

Comité local d'organisation
• Aassif Benassarou
• Sylvia Chalençon
• Hervé Deleau
• Eric Desjardin
• Didier Gillard
• Romain Guillemot
• Laurent Lucas
• Céline Loscos
• Jean-Michel Nourrit
• Nicolas Passat
• Stéphanie Prévost
• Yannick Remion
• Barbara Romaniuk
• Gilles Valette
• Philippe Vautrot

et nos thésards : Ludovic Blache, Exavérine Clin, Muhannad Ismaël et Tianatahina J-F. Randrianasoa

Comité de programme
• Pierre Alliez
• Olivier Aubreton
• Boulbaba Ben Amor
• Guillaume Boisson
• Gouenou Coatrieux
• David Coeurjolly
• Michel Couprie
• Florence Denis
• Benoit Huet
• Joel Jung
• Antoine Robert
• Xavier Rolland
• Neus Sabater
• Olivier Strauss
• Gérard Subsol
• Vincent Vidal
• Christian Wolf

Table des matières

3D

Visage / mouvement humain / suivi

Contribution des descripteurs de texture LBP à la classification d'images de dentelles

W. Ben Soltana, A. Porebski, N. Vandenbroucke, A. Ahmad et D. Hamad

Laboratoire d'Informatique Signal et Image de la Côte d'Opale
Maison de la Recherche Blaise Pascal 50, rue Ferdinand Buisson
BP 719, 62228 Calais Cedex France

Résumé
L'analyse d'image de dentelles présente un défit dans le domaine du traitement de l'image. Ceci est lié principalement à la nature complexe de la dentelle qui est généralement constituée de plusieurs parties avec des textures différentes : le fond, le motif, etc. Dans cet article, nous étudions séparément le comportement de trois descripteurs : l'histogramme d'image (HistI) et deux variantes des motifs binaires locaux (LBP) extraits des images de dentelles en présence du facteur de rotation. Ces variantes sont présentées par l'histogramme des LBP (LBP-B) et la transformée de Fourier appliquée sur les histogrammes de LBP (LBP-FFT). Par la suite, nous analysons l'apport de la fusion des données au niveau descripteur et au niveau score dans les différentes expérimentations. Le taux de classification évalue le degré de discrimination de chaque descripteur via le classifieur des plus proches voisins (k-ppv). Les résultats expérimentaux montrent qu'en l'absence de transformation, LBP-B, LBP-FFT et HistI fusionnés au niveau score génèrent la meilleure performance. En présence de changement de rotation, LBP-FFT et HistI fusionnés dans le même niveau produisent le meilleur taux de classification.

Abstract
The images of lace textile are particularly difficult to be analyzed in digital form using classical image processing techniques. The major reasons of this difficulty emerge from the complex nature of lace which generally has different textures in its constituents like the background and motives, etc. In this paper, we study separately the behavior of Image Histogram (HistI) and Local Binary Patterns (LBP) on image extracts of lace in presence of rotation. We further evaluate two variants of LBP ; primarily the hitogramme of LBP (LBP-B) and secondly the Applied Fourier Transform on the histogrammes of LBP (LBP-FFT). Consequently, we analyze the contribution of data fusion on feature level and on the score level in the different experimentations. The classification rate evaluates the discrimination degree of each descriptor via the k-ppv classifier. The experimental results indicate that the LBP-B, LBP-FFT and HistI combined at score level generate the best performance in absence of transformations. Whereas, LBP-FFT and HistI combined at the same level generate the best classification rate, in the presence of rotation.

Mots clé : Image de dentelle, analyse de texture, classification, LBP, FFT, invariance, fusion des données, k-ppv.

1. Introduction

La recherche d'images par le contenu consiste à retrouver, dans une base de données, des images visuellement similaires à une image requête. Généralement, le contenu des images (forme, texture, couleur...) est caractérisé par une signature (distribution, vecteur d'attributs...). Dans l'étape de recherche, une signature est extraite de l'image requête pour être comparée suivant différentes mesures de similarité à l'ensemble des signatures stockées. Dans ce cas, la classification garantit un gain de temps au niveau du processus de comparaison. Les images de la base, qui ont produit les plus grands scores de ressemblances, sont affichées et considérées comme les images les plus similaires. Plusieurs applications de recherche d'images par le contenu ont été développées sur les images de textures [OMP*02], de visages [CBF05], etc. Cependant, à notre connaissance, il n'existe pas d'application de recherche d'images de dentelles. Ceci peut être dû à la nature complexe de la dentelle. En effet, une dentelle est généralement constituée de plusieurs parties possédant des textures différentes : le fond, le motif, etc. De plus, la diversité des échantillons de dentelles et les conditions d'acquisition peuvent aussi rendre difficile leurs traitements.
Afin de décrire convenablement l'image de dentelle par des descripteurs invariants et discriminants, nous proposons

d'étudier séparément le comportement de trois types de descripteurs : l'histogramme d'image ($HistI$) et deux variantes des motifs binaires locaux (LBP) extraits des images de dentelles en présence du facteur de rotation. Ces variantes sont l'histogramme du LBP basique ($LBP-B$) et la transformée de Fourier appliquée sur les histogrammes de LBP ($LBP-FFT$). Afin d'exploiter les avantages de chaque descripteur, nous analysons ensuite l'apport de la fusion des données sur deux niveaux dans les différentes expérimentations. Le premier niveau, appelé niveau descripteur, opère avant l'opération de classification. Il consiste à concaténer les descripteurs pour obtenir un seul descripteur caractérisant l'image de dentelle. Le deuxième niveau, nommé niveau score, exploite les scores générés par les différents classifieurs propres à chaque descripteur.

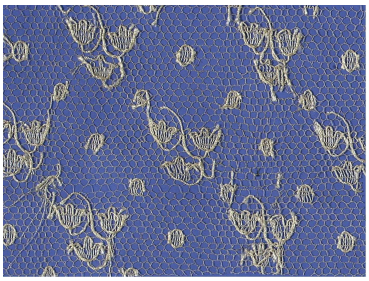
Cette étude a été effectuée dans le cadre du projet Interreg CRYSALIS impulsé par la Cité Internationale de la Dentelle et de la Mode de Calais (CIDM) pour la construction d'une tissuthèque de dentelles mécaniques. L'idée est de promouvoir la place de la filière textile dans des régions historiquement dédiées à ce domaine d'activité. Outre l'aspect culturel, le projet concerne la création d'une base de données d'images d'échantillons de dentelles. Dans ce cadre, nous avons élaboré une bibliothèque numérique avec outil de recherche en mode multicritères : langage contrôlé, langage libre ou par image requête de dentelle.

La suite de cet article est organisé comme suit : la section 2 présente les caractéristiques des dentelles à analyser. La section 3 expose les différents descripteurs de texture que nous avons testés. Les sections 4 et 5 décrivent respectivement la méthode de classification et la fusion d'information. La section 6 présente le protocole expérimental dont les résultats sont analysés dans la section 7. Finalement, la section 8 conclut cet article tout en présentant des perspectives.

2. Les images de dentelle

Les différentes images de dentelles appelées échantillons de dentelles ont été extraites à partir des registres conservés dans la cité internationale de la dentelle et de la mode. Les registres originaux sous forme de papier ont été scannés par un expert documentaliste qui a fourni des images en couleur de résolution 600 ppp. Les dentelles sont très variées suivant plusieurs critères comme le type des motifs (florale, animale,...), la nature des motifs (coton, textile,..), le type des fonds de dentelles et leurs couleurs (cf. figure 1.(a) et figure 1.(b)).

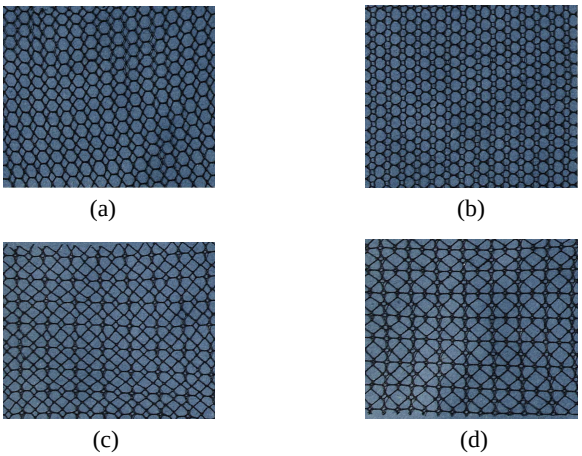
Cette diversité des dentelles possède néanmoins une spécificité commune. En effet, il existe un ensemble limité et bien défini de fonds de dentelles (figure 2). Dans cet article, nous proposons tout d'abord de reconnaître les différents types de dentelles en les classant selon leur fond. Cette application représente un véritable challenge. En effet, outre les conditions d'acquisition qui peuvent être variables, certaines classes de fonds sont très similaires et il est difficile, même pour l'oeil humain, de discriminer certaines classes. La figure 2 montre quatre exemples de fonds de dentelles où cette ressemblance est observable entre les fonds ((a),(b)) et ((c),(d)). La nature élastique des dentelles accentue également la difficulté de les identifier.



(a) Exemple de dentelle avec un fond blanc et des motifs

(b) Exemple de dentelle avec un fond noir et des motifs

Figure 1: *Exemple de dentelles conservées à la cité internationale de la dentelle et de la mode*



(a) (b) (c) (d)

Figure 2: *Exemple de fonds de dentelles*

3. Les descripteurs

Dans notre étude, trois descripteurs sont analysés. L'histogramme de l'image en niveau de gris est considéré comme un descripteur de référence. Les deux autres descripteurs sont choisis en se basant sur la spécificité des images de dentelles qui présentent des textures similaires. Ces descripteurs, qui ont prouvé leurs capacités discriminantes pour l'information de texture [OPH02], [OPM02], [AMP09], [GAMP12], sont l'histogramme des motifs binaires locaux et sa transformée de Fourier. Toutes les performances ont été évaluées séparément et aussi à la suite d'une opération de fusion.

3.1. Histogramme de l'image de dentelle

Dans notre travail, les images de dentelles sont des images couleurs. Elles ont été transformées en images en niveaux de gris. L'histogramme (*HistI*) est défini comme une fonction discrète qui associe à chaque valeur niveau de gris le nombre de pixels prenant cette valeur.

3.2. Histogramme des motifs binaires locaux

L'opérateur des motifs binaires locaux (*LBP*) a été proposé à la fin des années 90 par Ojala [OPH02], [OPM02]. L'idée de cet opérateur de texture est d'assigner à chaque pixel un code dépendant des niveaux de gris de son voisinage. Le niveau de gris du pixel central (i_c) de coordonnées (x_c,y_c) est comparé à celui de ses voisins (i_n) suivant l'équation (1) :

$$\begin{aligned} &LBP(x_c,y_c) = \textstyle\sum_{m=0}^{p-1} s(i_n - i_c) \times 2^n \\ &s(i_n - i_c) = 1 \quad si \quad i_n - i_c \geq 0 \\ &\qquad\qquad 0 \quad si \quad i_n - i_c < 0 \end{aligned} \tag{1}$$

avec p est le nombre de pixels voisins. Dans notre travail, nous considérons un voisinage de 3*3 d'où p=8 voisins. Nous obtenons donc, comme pour une image en niveaux de gris, une matrice contenant des valeurs des *LBP* comprises entre 0 et 255 pour chaque pixel . Un histogramme est calculé en se basant sur ces valeurs pour former le descripteur $LBP-B$.

3.3. Les coefficients de Fourier

Dans certains travaux [AMP09], [GAMP12], la transformée de Fourier rapide (FFT) est appliquée sur les histogrammes de *LBP* suivant l'équation (2) :

$$\begin{aligned} &H2(k) = \textstyle\sum_{i=1}^{N} h(i) \times exp^{(-j*2*\pi*(k-1)*(i-1)/N)}, \\ &1 \leq k \leq N \end{aligned} \tag{2}$$

avec h est l'histogramme des *LBP* et N représente sa dimension. Par la suite, l'amplitude des coefficients de la transformée de Fourier est retenue suivant l'équation (3) :

$$|H_{FFT}(k)| = \sqrt{H2(k) * \overline{H2(k)}} \tag{3}$$

avec $\overline{H2(k)}$ est le conjugué de $H2(k)$. Les nouvelles valeurs calculées permettent d'obtenir le descripteur $LBP-FFT$.

4. Classification d'images de dentelles

Après l'étape d'extraction des caractéristiques, nous procédons à la classification des images en se basant sur les différents descripteurs. Dans ce travail, la méthode des k plus proches voisins (*k-ppv*) [AHP04], [GAMP12] a été considérée vu son efficacité et sa flexibilité au niveau des protocoles d'expérimentation. Dans ce contexte, nous disposons d'une base de données d'apprentissage constituée de N couples images de classe connue. Pour estimer la classe d'une image I à classer, la méthode des (*k-ppv*) consiste à prendre en compte les k images d'apprentissage les plus proches de cette image selon une distance à définir. Dans notre problème de classification, nous retiendrons la classe qui possède la distance minimale avec l'image I. La distance L1 (eq. 4) a été retenue [AHP04], [AMP09], [GAMP12] pour calculer la distance entre deux vecteurs de descripteurs x et y.

$$d(x,y) = \sum_{j=1}^{B} |(x_j - y_j)| \tag{4}$$

avec x =(x_1,..,x_B) et y =(y_1,..,y_B) où B est la taille des vecteurs correspondants.

5. La fusion d'information

L'opération de la fusion d'information est une solution adoptée pour pallier les limitations imposées individuellement par les descripteurs et leurs classifieurs correspondants. En général, une opération de fusion permet d'avoir une décision finale plus précise en choisissant convenablement la méthode de fusion [SHA*10]. Elle s'applique avant ou après l'opération de classification. Nous distinguons deux grandes catégories [JNR05] : la fusion au niveau des descripteurs et la fusion au niveau de scores [JDM00], [SHA*11].
Dans la première catégorie, il s'agit de concaténer les vecteurs de descripteurs pour les préparer à l'opération de classification. Pour cela, une étape de normalisation s'avère primordiale afin que les données des descripteurs soient définies dans le même intervalle. Dans notre travail, la normalisation par la norme 1 (eq. 5) a été considérée.

$$x' = \frac{x}{\|x\|_1} \tag{5}$$

où x est un vecteur de caractéristique et x' est le vecteur normalisé.
Dans la catégorie de fusion au niveau score, chaque classifieur propre à un descripteur opère indépendamment, ensuite les réponses sous forme de décisions ou de distances sont fusionnées. Ici, chaque classifieur reçoit en entrée un vecteur de caractéristique x et génère un ensemble de distances $e_j(x)$ (eq.6) propres aux différents prototypes ou classes de la base d'apprentissage.

$$e_j(x) = [d_j^1, d_j^i, .., d_j^C] \tag{6}$$

où d_j^i est la distance attribuée par le classifieur j à la classe i, et, C est le nombre de classe dans la base d'apprentissage. Toutefois, pour une opération de fusion, ces mesures ne sont pas toujours dans le même intervalle, une normalisation s'avère donc nécessaire [SHA*10] [HLS94]. Dans notre travail, la normalisation max-min (eq.7) a été appliquée sur les éléments du vecteur e_j pour produire un nouveau vecteur de distances e_{2j} =[d_{2j}^1,d_{2j}^i,..,d_{2j}^C] avec des éléments d_{2j}^i où :

$$d_{2j}^i = \frac{d_j^i - e_{j,min}}{e_{j,max} - e_{j,min}} \tag{7}$$

avec $e_{j,max}$ et $e_{j,min}$ représentent respectivement la valeur maximale et la valeur minimale du vecteur e_j.
Par la suite, pour chaque prototype dans la base d'apprentissage, nous calculons une distance S_i selon une règle précise. En tenant compte de la simplicité et de l'efficacité

[KHDM98] [SHA*11] de la règle de somme simple, nous l'adoptons pour le calcul de S_i suivant l'équation (8) :

$$S_i = \sum_{j=1}^{K} d_{2i}^{j} \quad (8)$$

où K est le nombre de classifieurs. Une règle de décision (eq. 9) pour chaque image requête I consiste alors à choisir la classe C_i pour laquelle la distance S_i est la plus petite avec :

$$i = argmin_{i=1,..M}(S_i) \quad (9)$$

6. Protocole expérimental

Le protocole expérimental permet de présenter la décomposition de la base d'images pour les différentes expérimentations. Initialement, la base d'images est constituée de 492 images de dentelles obtenues à partir des 41 fonds de dentelles. Chaque fond de dentelles a généré 12 imagettes de fonds de dentelles de taille fixe (150*150 pixels).
Cette base a été décomposée en deux sous-bases : base d'apprentissage et base de test (décomposition 1). La base de test a ensuite été enrichie par de nouvelles images de test correspondant aux images de test de départ auxquelles nous avons appliqué des rotations de 90°, 180° et 270° (décomposition 2). Toutes les images de la base se présentent avec la même échelle. Le tableau 1 affiche les décompositions citées. La décomposition 1 a permis de tester les images sans transformation. La décomposition 2 permet d'évaluer l'impact du changement de rotation sur la performance des descripteurs.

Table 1: *Les différentes décompositions de la base d'images*

Numéro de la décomposition	Taille de la base d'apprentissage	Taille de la base de test
1	246 (6*41)	246 (6*41)
2	246 (6*41)	984 (4*6*41)

Le taux de classification est adopté comme critère d'évaluation de nos descripteurs et de nos schémas de fusion. Il représente le nombre d'images test correctement classées sur le nombre total d'images test. Pour chaque décomposition, nous avons appliqué une validation croisée comptant 20 expériences. Dans chaque expérience, nous avons sélectionné aléatoirement pour chaque image de fond de dentelle 6 images d'apprentissage et 6 images de test. Nous adoptons par la suite le taux de classification moyen pour exprimer la performance de chaque descripteur. Nous notons la fusion au niveau descripteur et au niveau score respectivement par Fusion I et Fusion II.

7. Résultats expérimentaux

Les résultats expérimentaux ont permis d'observer le comportement des trois descripteurs par rapport aux transformations appliquées.

7.1. Résultats obtenus sans transformation

Les images de la décomposition 1 ont permis de tester tous les descripteurs ainsi que les différents schémas de fusion. Dans les deux niveaux de fusion, nous avons évalué toutes les combinaisons possibles de fusion des trois descripteurs ($HistI$, $LBP-B$, $LBP-FFT$). Le tableau 2 présente les différents taux de classification. Les trois premières lignes affichent les taux de classification individuels des trois descripteurs. Nous détectons une légère chute de performance pour le descripteur $LBP-FFT$ contrairement aux autres descripteurs dans la Fusion I à cause de l'opération de normalisation appliquée. Nous observons que $LBP-B$ a généré un résultat très encourageant avec un taux de classification de l'ordre de 97.05% suivi par la performance de $LBP-FFT$ avec 86.98% dans Fusion I et 87.56% dans Fusion II. Il est clair que les motifs binaires locaux ont permis d'obtenir un gain de l'ordre de 24% par rapport au descripteur histogramme de l'image. Ceci montre que l'information de la texture est plus discriminante que l'information contenue en analysant uniquement les niveaux de gris des pixels sans leur interaction spatiale.
Dans la fusion au niveau descripteur, aucun schéma de fusion n'a permis d'améliorer la performance atteinte par le $LBP-B$. Ceci est dû aux faibles performances obtenues par $HistI$ et $LBP-FFT$ comparées à la performance du $LBP-B$. Par contre, la concaténation des deux descripteurs $HistI$ et $LBP-FFT$ a permis d'améliorer le taux de classification de 13% par rapport à la performance individuelle du descripteur $HistI$ et 1 % par rapport à la performance individuelle du descripteur $LBP-FFT$. Cette amélioration met l'accent sur la complémentarité entre ces deux descripteurs.

	Fusion I	Fusion II
HistI	0.7348	0.7348
LBP-B	0.9705	**0.9705**
LBP-FFT	0.8689	0.8756
LBP-B + HistI	0.9163	0.9657
LBP-B + LBP-FFT	0.9669	0.9648
HistI + LBP-FFT	0.8774	0.9246
LBP-B + HistI + LBP-FFT	0.9443	**0.9732**

Table 2: *Taux de classification moyen dans les deux niveaux de fusion avec la décomposition 1*

Dans la fusion au niveau de scores, toutes les performances (Fusion II) sont présentées dans le tableau 2. D'une part, nous remarquons que le seul schéma de fusion qui combine les trois descripteurs a permis d'améliorer légèrement la performance du descripteur $LBP-B$. Dans ce cas, nous notons une amélioration de l'ordre de 0.27% en taux de classification. Cette faible augmentation peut être liée en partie à la règle de somme simple (eq. 8) qui n'est pas pondérée. Par conséquence, elle accorde la même confiance et importance à tous les descripteurs qui ont des performances distinctes.

7.2. Résultats obtenus avec transformation de rotation

Les images de la décomposition 2 ont permis d'évaluer tous les descripteurs individuellement ainsi que les schémas de fusion par rapport à l'invariance en rotation. Le tableau

3 expose les différents résultats. L'histogramme d'image et $LBP-FFT$ étant des descripteurs invariants à la rotation, les résultats de classification obtenus sont similaires à ceux obtenus avec les images de la décomposition 1. Une comparaison avec les travaux de l'état de l'art est difficile puisque les bases d'images utilisées sont différentes. Cependant nos résultats confortent ceux obtenus par [AMP09] et [GAMP12] concernant la comparaison des attributs invariants en rotation par rapport aux attributs basiques. D'après ce tableau, nous remarquons l'impact négatif de cette transformation sur la performance du descripteur $LBP-B$. Une chute de 49% du taux de classification est observée par rapport au scénario 1. Pour diminuer cette influence négative, nous procédons à la fusion des descripteurs. Dans la fusion I, la concaténation de tous les descripteurs a permis d'atteindre 85.01% de taux de classification. Il est clair que le fait d'intégrer $LBP-B$ dans la stratégie de fusion ne permet pas d'améliorer la performance globale par rapport au meilleur taux de classification généré (86.89%) par $LBP-FFT$. Dans cas, la non sélection de ce descripteur dans le schéma de fusion ($HistI$ + $LBP-FFT$) assure d'atteindre une meilleure performance globale de l'ordre de 87.74% de taux de classification.

	Fusion I	Fusion II
HistI	0.7348	0.7348
LBP-B	0.4731	0.4731
LBP-FFT	0.8689	0.8756
LBP-B + HistI	0.7172	0.5943
LBP-B + LBP-FFT	0.5745	0.5758
HistI + LBP-FFT	0.8774	**0.9246**
LBP-B + HistI + LBP-FFT	0.8501	0.7991

Table 3: *Taux de classification moyen dans les deux niveaux de fusion avec la décomposition 2*

Dans la fusion II, le même comportement est observée pour le descripteur $LBP-B$. En effet, le seul schéma de fusion capable de donner la meilleur performance (92.46%) résulte de la fusion des scores générés par les classifieurs propres à $LBP-FFT$ et $HistI$. Ce résultat approuve l'avantage de la fusion au niveau de score par rapport à la fusion au niveau descripteur vu dans les deux scénarios d'expérimentation.

8. Conclusion

Dans cet article, nous avons analysé le comportement de trois descripteurs que sont l'histogramme d'image en niveau de gris et les motifs binaires locaux ($LBP-B$ et $LBP-FFT$) extraits des images de dentelles en présence du facteur de rotation. Le but de cette étude est d'obtenir un descripteur discriminant pour la classification des images de dentelles. Par la suite, nous avons analysé l'apport de la fusion au niveau descripteur et au niveau score dans les différentes expérimentations. Les résultats montrent qu'en l'absence de transformation, $LBP-B$, $LBP-FFT$ et $HistI$ fusionnés au niveau score génèrent la meilleure performance de l'ordre de 97.32% d'images bien classées. Dans le cas de présence de changement de rotation, la combinaison des deux descripteurs $LBP-FFT$ et $HistI$ dans le même niveau de fusion donne le meilleur taux de classification avec 92.46% d'images bien classées.

Comme perspectives, nous envisagerons d'étendre les expériences avec le LBP multi-résolution en modifiant les paramètres de rayon (R) et de voisinage (P) et en utilisant d'autres types de LBP comme LBP uniforme et LBP invariant à la rotation. Nous intégrerons aussi les changements d'échelle comme une nouvelle transformation. En plus, nous étudierons l'apport d'autres méthodes de fusion par rapport aux différentes transformations.

9. Remerciements

Ce travail est effectué au sein du projet Interreg IV A, 2 Mers CRYSALIS.

Références

[AHP04] AHONEN T., HADID A., PIETIKÂINEN M. : Face recognition with local binary pattern. *Computer Vision, ECCV* (2004), 469–481.

[AMP09] AHONEN T., MATAS J., PIETIKÂINEN M. : Rotation invariant image description with local binary pattern histogram fourier features. *Image Analysis* (2009), 61–70.

[CBF05] CHANG K. I., BOWYER K. W., FLYNN P. J. : An evaluation of multimodal 2d+3d face biometrics. *IEEE Trans. Pattern Anal. Mach. Intell. Vol. 27*, Num. 4 (2005), 619–624.

[GAMP12] GUOYING Z., AHONEN T., MATAS J., PIETIKÂINEN M. : Rotation-invariant image and video description with local binary pattern features. *IEEE Transactions on Image Processing. Vol. 21*, Num. 4 (2012), 1465–1477.

[HLS94] HUANG Y. S., LIU K., SUEN C. Y. : The combination of multiple classifiers by neural network approach. *Journal of Pattern Recognition and Artificial Intelligence. Vol. 9*, Num. 3 (1994), 579–597.

[JDM00] JAIN A. K., DUIN R. P. W., MAO J. : Statistical pattern recognition : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 22*, Num. 1 (2000), 4–37.

[JNR05] JAIN A., NANDAKUMAR K., ROSS A. : Score normalization in multimodal biometric systems. *Pattern Recognition. Vol. 38*, Num. 12 (2005), 2270–2285.

[KHDM98] KITTLER J., HATEF M., DUIN R. P. W., MATAS J. : On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 20*, Num. 3 (1998), 226–239.

[OMP*02] OJALA T., MÂENPÂÂ T., PIETIKÂINEN M., VIERTOLA J., KYLLÔNEN J., HUOVINEN S. : Outex new framework for empirical evaluation of texture analysis algorithms. *In Proceedings of the 16th International Conference on Pattern Recognition. Vol. 1* (2002), 701–706.

[OPH02] OJALA T., PIETIKÂINEN M., HARWOOD D. :

A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition. Vol. 29*, Num. 1 (2002), 51–59.

[OPM02] OJALA T., PIETIKÂINEN M., MÂENPÂÂ T. : Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Recognition. Vol. 24*, Num. 7 (2002), 971–987.

[SHA*10] SOLTANA W. B., HUANG D., ARDABILIAN M., CHEN L., AMAR C. B. : Comparison of 2d/3d features and their adaptive score level fusion for 3d face recognition. *3D Data Processing, Visualization and Transmission (3DPVT), Paris* (May 2010), 17–20.

[SHA*11] SOLTANA W. B., HUANG D., ARDABILIAN M., CHEN L., AMAR C. B. : A mixture of gated experts optimized using simulated annealing for 3d face recognition. *International Conference on Image Processing (ICIP), Brussels* (Septembre 2011), 11–14.

Un jeu, des images, des clics et du texte : collecte implicite de données visuelles et sémantiques

Axel Carlier [1] et Vincent Charvillat[1]

[1]Université de Toulouse, IRIT-ENSEEIHT, 2 rue Camichel, 31000 Toulouse
axel.carlier@enseeiht.fr vincent.charvillat@enseeiht.fr

Résumé
Nous décrivons un corpus de données visuelles et sémantiques collectées à partir d'un jeu. Nous avons conçu ce jeu pour deux joueurs qui coopèrent à distance sur le Web. Les données collectées sont directement utilisables pour résoudre des problèmes de vision (par exemple des problèmes de détection, de segmentation, d'étiquetage sémantique). La collecte est toutefois implicite au sens où le jeu n'a pas pour but explicite de détecter ou segmenter des objets présents dans une image. Le corpus inclut plus de 3,250 jeux basés sur 104 images et comporte des annotations textuelles et des données spatiales (clics, relations spatiales). Dans cet article, nous expliquons pourquoi et comment utiliser ces données pour différentes applications visant la compréhension des images. Nous montrons surtout qu'elles sont suffisamment riches pour superviser, dans un sens à définir, une analyse sémantique globale du contenu visuel. Le corpus est rendu accessible aux chercheurs.

Mots clé : Jeu GWAP, Corpus de données, Sémantique, Détection, Segmentation

1. Introduction

L'analyse sémantique d'images au sens le plus ambitieux est parfois nommée *holistic image understanding* ou *image parsing* dans la littérature anglo-saxonne. Il s'agit d'un des problèmes centraux en vision par ordinateur et sans doute aussi d'un des plus ardus. Il s'agit d'associer une étiquette sémantique à chaque pixel pour assurer l'interprétation complète d'une image. Cela est si compliqué que les chercheurs préfèrent souvent décomposer le problème en plusieurs sous-problèmes comme la classification d'images, la détection d'objets ou la segmentation. Ces sous-problèmes, bien que très fouillés ces dernières années, restent ouverts. A titre d'exemple et malgré plus de vingt ans de travaux dans ce domaine, le problème de la segmentation d'un objet d'intérêt n'est pas encore résolu par des approches générales et automatiques.

Les recherches dans ce domaine se scindent en deux catégories principales :

– *Les approches par apprentissage artificiel* : la résolution de problèmes de segmentation s'appuie dans ce cas sur des bases d'apprentissage de grande taille qui regroupent des vérités terrains, c'est-à-dire des masques de segmentation considérés comme exacts pour différents objets. A partir de ces données d'apprentissage, des prédicteurs sont appris pour segmenter, par inférence, de nouvelles images. Bien qu'efficaces pour certains objets, ces approches se heurtent à des difficultés pour différentes catégories d'objets. Par exemple, le vainqueur de la dernière compétition *PASCAL VOC Challenge, 2012* pour la segmentation obtient des scores modestes dans le cas des bicyclettes, des chaises, des tables ou autres canapés.
– *Les approches interactives de la segmentation* : on admet dans ce cas que seule une segmentation semi-automatique est atteignable. Ce qui revient à dire que le *gap sémantique* est délicat à combler sans intervention humaine. Placer un (ou des) utilisateur(s) dans le processus permet d'initialiser, de superviser, de corriger des algorithmes de segmentation au travers d'interfaces adaptées.

Dans les deux cas précédents, notons bien qu'une supervision humaine est présente, ou bien dans la constitution de la vérité terrain ou bien au travers d'une segmentation interactive. La question du recrutement de ces humains, en particulier lorsqu'ils servent d'experts, est ouverte. Les approches dites de *crowdsourcing* donnent une réponse au travers de plateformes comme *amazon mechanical turk* ou *microworkers*. Selon Luis Von Ahn qui est parmi les premiers à avoir parlé d'*human computation*, une alternative est possible au travers de jeux appelés GWAP *Games With A Purpose*. Alors qu'un paiement récompense les humains travaillant sur *amazon mechanical turk*, la motivation pour participer à un jeu est ou devrait être naturelle. Par principe, un jeu GWAP doit simultanément être amusant et pouvoir contribuer à la ré-

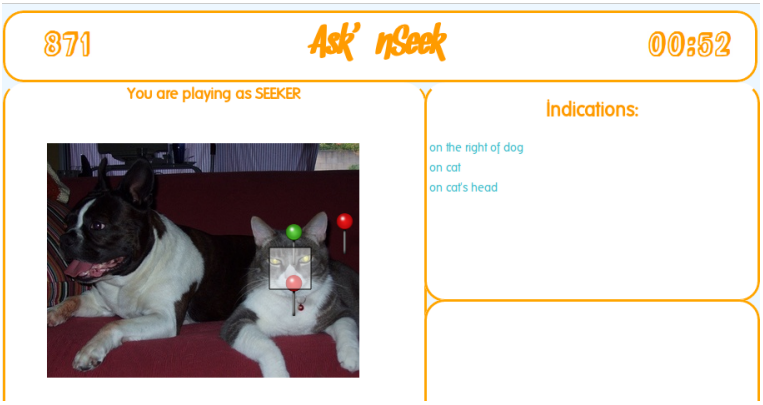


Figure 1: *Ecran du jeu de l'enquêteur qui demande des indices relatifs au chien, au chat et finalement à la tête du chat.*

solution d'un problème. Un jeu GWAP parmi les plus populaires est le jeu ESP [VAD04] qui réunit deux joueurs à distance à travers une application web. Les joueurs gagnent des points lorsqu'ils annotent les mêmes images avec des étiquettes similaires. Le scénario permet, statistiquement, d'annoter globalement les images utilisées par ESP.

Dans cet article, nous suivons la même voie en utilisant un jeu que nous avons proposé il y a deux ans [CMC12, SCGiN*13]. Cet article dresse un bilan des données que nous avons collectées après 3,250 jeux basés sur 104 images. Dans les paragraphes suivants, nous montrons le très fort potentiel du corpus de données visuelles et textuelles obtenu. Ce corpus sera publiquement accessible.

2. Corpus de données existants

De nombreux corpus de données sont disponibles pour évaluer les algorithmes de segmentation en vision par ordinateur. Parmi eux, deux jeux de données sont particulièrement populaires : le corpus de Berkeley *Berkeley Segmentation Dataset* [MFTM01] et le corpus du réseau d'excellence PASCAL *PASCAL VOC Dataset* [EVGW*10].

Celui issu de Berkeley (BSDS300 en résumé) est composé de 300 images auxquelles sont associées des segmentations manuelles opérées par des experts. Il y a plusieurs (typiquement entre 5 et 10) segmentations par image et chaque segmentation, pour une image donnée, est différente et donc complémentaire des autres. Ensemble ces segmentations fournissent une bon approximation des différentes interprétations possibles d'une même image. Notons que ce corpus a été étendu à 500 images. Le corpus PASCAL n'a pas uniquement été créé pour la segmentation mais aussi pour la détection d'objets, la classification d'actions, les détections de parties du corps etc. En 2012, à la compétition PASCAL, on pouvait dénombrer dans le corpus 11530 images associées à 27450 annotations liées à des régions d'intérêt (ROI) et 6929 segmentations. Le corpus *MSRA dataset* [LYS*11] est aussi reconnu dans le domaine de la détection d'objets. Ce corpus intègre 20840 images associées à une vérité terrain constituée de boîtes englobantes indiquant la localisation des objets les plus saillants dans chaque image.

Ces trois corpus sont intéressants car ils intègrent une grande quantité de données et une vérité terrain pour différents problèmes. Le corpus que nous présentons dans ce travail est de nature différente. Il intègre des images et des traces collectées au travers d'un jeu. Ces traces ne correspondent pas directement à des solutions aux problèmes de détection ou de segmentation auxquels on s'intéresse. Les traces collectées sont plus proches en nature des gribouillages (ou *scribbles*) utilisés pour apporter une supervision partielle des algorithmes de segmentation interactive.

La segmentation interactive a été beaucoup étudiée dans les dernières années et plusieurs corpus ont été proposés pour évaluer ces algorithmes. Le corpus *Grabcut* [RKB04] est bâti sur 50 images (dont 20 proviennent du corpus de Berkeley) et sur trois types de masques pour la vérité terrain. Gulshan et al [GRC*10] ont augmenté ce corpus avec des images de PASCAL pour atteindre un total de 151 images avec vérité terrain. McGuinness et al [MO10] proposent quant à eux un corpus de 96 images extraites de Berkeley auxquelles ils ajoutent des masques déterminés manuellement en guise de vérité terrain. Ce corpus est utilisé comme *benchmark* pour comparer différents algorithmes de segmentation interactive. Bien que plus proches de notre travail, ces corpus n'intègrent pas de traces utilisateurs comme notre ensemble de données. Une telle collecte a, par contre, été proposée dans d'autres travaux autour des GWAP.

Von Ahn et al ont par exemple publié un jeu de données contenant 100,000 images et des annotations provenant du jeu GWAP nommé ESP [VAD04]. Ce corpus est pratique pour tester des requêtes et des algorithmes de recherche d'images dans un contexte d'indexation textuelle. Il n'est pas naturellement adapté à l'évaluation d'algorithmes de détection ou de segmentation. Russell et al [RTMF08] ont aussi rendu public le corpus *LabelMe* provenant du jeu du même nom. Ce corpus intègre des tracés de polygones détourant les objets d'intérêt. Ce détourage est trop grossier pour évaluer des algorithmes de segmentation. A la différence du jeu que nous défendons, les joueurs ont clairement conscience de détourer ou segmenter les objets d'intérêt. La conception de notre jeu est telle que les joueurs n'ont pas conscience que leurs actions (mises en commun) contribuent indirectement à l'identification des objets visibles dans une image.

3. Le jeu

Le jeu que nous utilisons est un jeu à 2 joueurs déployé sur Internet. Il a été initialement introduit par Carlier et al. dans [CMC12]. Le jeu est non symétrique : le premier joueur appelé maître commence par cacher une cible (ou région cachée) sur une image, puis doit aider le second joueur appelé enquêteur (ou *seeker* dans la version anglaise) à la découvrir. Sur son écran, l'enquêteur voit la même image que le maître mais ne connait évidemment pas la position de la cible. Son but est de la trouver en cliquant en son sein.

Pour ce faire, l'enquêteur peut demander des indications au maître. Plus spécifiquement, l'enquêteur peut taper des nom d'objets qui apparaissent à l'écran c'est-à-dire dans l'image, puis demander au maître de localiser la cible par rapport à ces objets. Le maître peut répondre de 7 manières différentes. La cible peut se situer "à gauche", "à droite", "au dessus" ou "en dessous" de l'objet cité. Elle peut également être située sur l'objet, ou partiellement sur l'objet. Enfin, il arrive qu'il ne soit pas possible de catégoriser la position de la cible par rapport à l'objet avec l'une de ces 6 possibilités : dans ce cas la cible ne peut pas être reliée à l'objet. Grâce à ces indications, l'enquêteur va peu à peu réduire le champ des positions possibles pour la cible.

Ce mécanisme est illustré sur la figure 1. Dans la partie jouée sur cet exemple, l'enquêteur a commencé par demander au maître un indice lié au chien. Le maître a répondu que la cible était située à droite du chien. L'enquêteur a obtenu le droit de cliquer sur l'image (logiquement à droite du chien), mais a manqué la cible. Il a donc eu le droit de demander un second indice, et après avoir tapé le mot "chat", il a reçu l'indication que la cible était située "sur" le chat. Après avoir à nouveau manqué la cible, il a finalement obtenu l'information que la cible était située "sur la tête du chat" et a réussi à cliquer dans la région cachée, déclenchant du même temps l'apparition de la cible sur son image et la fin de la partie.

Comme nous allons l'expliquer avec plus de détails, les informations collectées sont directement utilisables pour résoudre des problèmes de vision (par exemple des problèmes de détection ou de segmentation). Pour autant, la collecte est implicite au sens où le jeu n'a pas pour but de détecter ou segmenter des objets présents dans une image.

4. Nature des données

Les données collectées via notre jeu sont produites par les interactions entre deux joueurs ayant les rôles de maître et d'enquêteur. Il est important de comprendre que les données produites par ces deux rôles sont distinctes et sont complémentaires en vue de la compréhension des images. L'enquêteur fournit deux types d'informations : des informations textuelles (lorsqu'il demande des indices) et des informations spatiales quand il clique sur l'image en espérant atteindre la région cachée. Le maître fournit deux autres types d'informations. Le premier type correspond à la position de la région qu'il sélectionne pour être découverte durant le jeu (dont on rappelle qu'il répond à un scénario de coopération). Le second type relève des relations spatiales qu'ils fournit entre les objets/indices demandés par l'enquêteur et la région cachée. Puisque l'enquêteur prend en compte ces informations avant de cliquer, ces relations spatiales sont aussi révélatrices de relations entre les objets et les clics produits. Nous détaillons maintenant les données collectées.

4.1. Informations textuelles

Les données textuelles (ou *tags*) saisis par l'enquêteur sont de formes variées. Comme la durée du jeu est limitée, les joueurs se contentent généralement de textes courts. La plupart des données sont réduites à un mot faisant référence à un objet présent dans l'image ou à une partie d'un tel objet. La partie droite de figure 2 illustre cela avec mes mots "tree", "plant" et "tiger" (mot) et "face", "tail", ou "tongue" qui sont des méronymes de "tiger". L'analyse (statistique) de ces données permet un étiquetage sémantique de l'image.

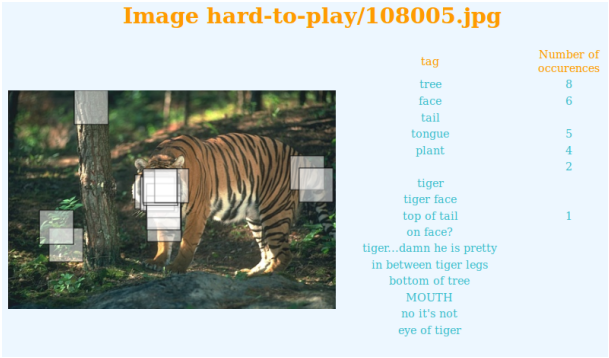


Figure 2: *Illustration compacte des jeux joués sur une image. A gauche, les régions cachées et à droite les textes et leur nombre d'occurrences.*

Il arrive aussi que les textes soient des groupes de mots qui peuvent être très informatifs en connectant un objet et une de ses parties ("eye of tiger", "tiger face") ou plus compliqués à interpréter ("in between tiger legs"). L'analyse de ces traces peut conduire à la détermination d'une hiérarchie d'objets au sens sémantique/ontologique.

Dans le cas des instances multiples d'objets, les traces collectées sont riches et complexes. Les textes permettent aux joueurs de distinguer différents objets de même catégorie.

Dans la figure 3, le texte "soldier in the foreground on the left" suggère qu'il y a des soldats à l'avant et à l'arrière plan. Nous avons observé, dans le corpus, une intéressante corrélation entre la longueur des textes et la présence d'instances multiples d'objets : pour gérer les ambiguïtés les textes se complexifient.

De manière évidente enfin, les données textuelles présentes dans notre corpus sont adaptées à la catégorisation d'images et à l'étude des co-occurrences sémantiques d'objets dans certaines catégories d'images.

4.2. Régions cachées

Les régions cachées sont les cibles positionnées par le maître du jeu. La figure 4 présente, dans la ligne centrale, toutes les régions cachées sur deux images du corpus par les joueurs maîtres. Il est intéressant de noter que les positions

soldier in the foreground on
the left
the left soldiers back
soldier on the right
blue berret
man on left
man to the right

Figure 3: *Une image, en haut, où des instances multiples conduisent à des textes visant à les distinguer, en bas.*

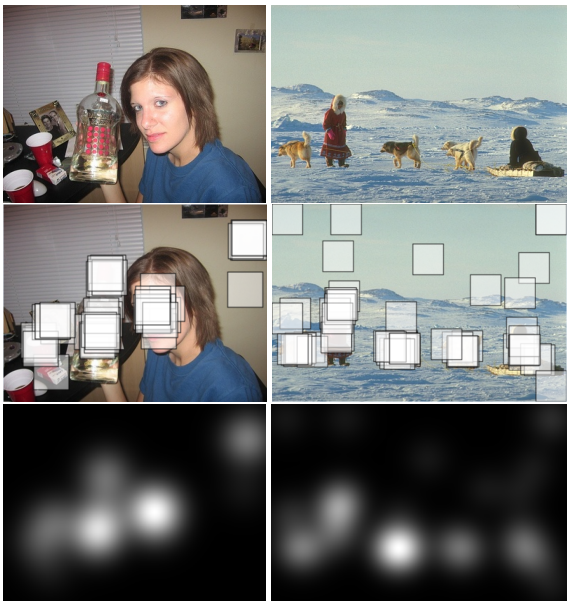


Figure 4: *Cartes de saillance issues des régions cachées par les joueurs pour deux images (chacune sur une colonne).*

de ces carrés ne sont pas aléatoires mais souvent placées sur les objets les plus importants, les plus discriminants et dans un sens, les plus saillants. Cela peut être expliqué par la nature coopérative du jeu : pour vite gagner des points, le maître du jeu doit placer la région cachée à un emplacement qui sera facilement identifié par l'enquêteur et cela grâce à quelques indices seulement. Au travers du corpus, les positions des régions cachées sont précurseurs de nouvelles cartes de saillance que nous illustrons en mélangeant des gaussiennes centrées sur les positions en question en bas de la figure 4. Ce résultat prometteur mérite une étude ultérieure et une comparaison avec d'autres mécanismes attentionnels que nous estimons possible grâce à la mise à disposition de notre corpus de données.

4.3. Clics et relations spatiales

Tous les clics des enquêteurs sont associés à un *tag* et à des relations spatiales (*above*, *below*,*on the left of*,*on the right of*,*on*,*partially on*).

Au dessus, En dessous, A gauche, A droite. La figure 5 illustre comment nous pouvons utiliser ces clics pour détecter des objets. Les points rouges sont les clics collectés au dessus du chat alors que les points bleus sont les clics collectés à droite du chat. Les lignes correspondent aux boîtes englobantes que nous pouvons utiliser pour détecter le chat si nous faisons entière confiance aux traces présentes dans le corpus (lignes en pointillés) ou si nous choisissons de résister aux erreurs potentielles (lignes pleines). Dans ce dernier cas, l'utilisation d'estimateurs de position robustes aux données aberrantes est nécessaire ; le plus simple d'entre eux est la médiane.

Sur. La figure 6 montre tous les clics de type "on" (sur un objet) obtenus pour une image. Il est clair que la distribution de ces points peut permettre de segmenter (ou de contraindre la segmentation de) certains des objets présents (hut, man,

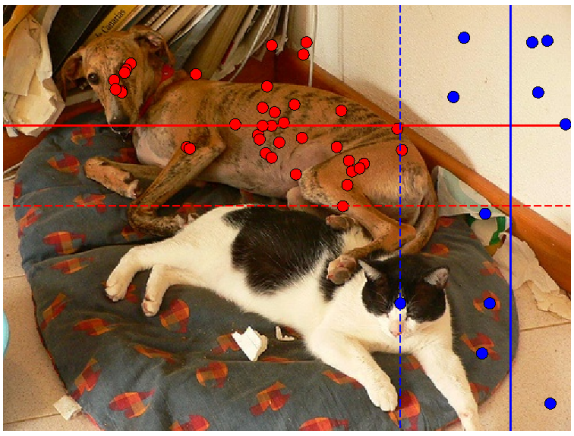


Figure 5: *Clics au dessus (en rouge) et à droits (en bleu) du chat. Le* tag *'cat' étant avec 'dog' le plus fréquemment collecté pour cette image.*

trees, stick). L'image en dessous montre les points de type "on" obtenus sur des parties (head, skirt, butt, feet) d'un objet qui les inclut (man).

Partiellement sur.. La figure 7 montre tous les clics situés "partiellement sur" l'objet "urn". Dans la version actuellement en ligne du jeu cette relation, assez mal comprise de certains joueurs, est remplacée par "sur le bord de". La figure montre cependant que les "partially on points" sont situés majoritairement au voisinage des contours de l'urne. Ce constat est vérifié sur beaucoup d'images du corpus. Ces données sont une source intéressante d'information pour la segmentation en conjonction avec les points cliqués "sur" un objet et ceux situés en dehors de l'objet (clics "above", "below", "left" et "right").

Figure 6: *Les clics sur des objets ou* 'on clicks'.

Les traces collectées via notre jeu sont à la fois variées et complémentaires les unes par rapport aux autres. Nous pouvons les utiliser pour la segmentation d'un objet d'intérêt présent dans une image mais aussi, en généralisant à plusieurs objets, pour la compréhension globale de la scène représentée par une image. La figure 6 illustre ce potentiel vis-à-vis d'un objectif d' **image parsing**. L'originalité du corpus est aussi de reposer sur des relations spatiales nombreuses qui structurent les informations textuelles, les *tags* et les clics. Des relations spatiales entre objets, entre instances multiples de mêmes objets ou entre parties d'objets peuvent être extraites du corpus.

Figure 7: *En vert, tous les points qui ont été cliqués comme étant "partiellement sur" l'urne.*

5. Collecte des données

Dans ce paragraphe nous expliquons comment les traces ont été collectées pour former le corpus présenté. D'abord nous avons choisi des images déjà sélectionnées par McGuinness et al [MO10] pour leur propre corpus. Nous sommes partis de cet ensemble car il avait déjà été utilisé pour comparer différentes approches de segmentation interactive. En plus des 96 images de McGuinness et al. nous avons aussi sélectionné 8 images de PASCAL pour être jouées intensément. Les traces sur ces images peuvent être utilisées pour déterminer combien de jeux sont suffisants pour obtenir des résultats satisfaisants étant donné un problème à résoudre (détection, segmentation etc.).

Le corpus intègre des images de difficultés différentes vis-à-vis du jeu. Des images trop simples (intégrant un seul objet centré de grande taille par rapport à celle de la région à cacher) se révèlent difficiles à utiliser durant le jeu par manque d'indices visuels par rapports auxquels l'enquêteur peut guider sa recherche. Inversement, des images plus complexes intégrant plusieurs objets se révèlent plus simples à manier au travers du jeu. Lorsqu'un nouveau joueur se connecte à la plateforme de jeu, des images faciles à jouer lui sont d'abord proposées.

Dans tous les cas, les joueurs débutants doivent suivre un tutoriel vidéo et deux tutoriels interactifs les initiant aux rôles de maître ou d'enquêteur respectivement. Aucune explication n'est donnée quant à l'intérêt du jeu vis-à-vis de la recherche scientifique et de la compréhension semi-automatique d'images. Par contre, des recommandations sont fournies pour rappeler le caractère coopératif du jeu, pour indiquer que des *tags* simples permettent de bien gérer la limite de temps etc. Comme le jeu n'est pas célèbre, nous avons lancé plusieurs campagnes auprès d'étudiants et

de réseaux sociaux. Les campagnes ont impliqué plus de 50 joueurs qui se sont réunis à des heures précises pour faciliter l'appariement aléatoire de joueurs en ligne.

6. Le corpus en pratique

Caractéristiques du corpus : Les données des milliers de jeux joués ont été stockés dans une base de données MySQL. Nous rendons publique la partie scientifiquement pertinente de cette base de données. La base de données est constituée des tables suivantes :

- *User*. Cette table contient les *login* et les *password* cryptés des joueurs. Ces informations ne seront pas rendues publiques. Du point de vue scientifique, les seules informations intéressantes de cette table sont l'âge et le genre des joueurs.
- *Game*. Pour chaque jeu, nous stockons l'image qui a été jouée, les identificateurs (*ID*) des joueurs formant une paire maître-enquêteur, le score final, le temps restant à la fin du jeu. On peut ainsi simplement déduire de ce dernier attribut le temps effectif du jeu puisque le chronomètre interrompt le jeu après 120 secondes.
- *Region*. Pour chaque jeu, nous stockons la région cachée par le maître. La région est identifiée par ses coordonnées centrales avec, de plus, sa hauteur et sa largeur. Les figures 2 et 11 montrent toutes les régions cachées sur des images particulières.
- *Indication*. Cette table est la plus intéressante puisqu'elle contient les textes (*tags*) soumis par les enquêteurs, les relations données en retour par les maîtres et les coordonnées des clics qui s'en suivent. Cet ensemble forme une indication. Chaque ligne de la table du même nom consiste en une indication pour un jeu donné.

Interface : Nous fournissons une interface web pour visualiser les traces. Cette interface permet, pour chaque image, d'afficher rapidement les positions des régions cachées durant l'ensemble des jeux concernant cette image. En même temps, les informations textuelles utilisées pour décrire le contenu de l'image sont aussi restituées. Une capture de cette interface est visible dans la figure 2. L'interface peut aussi être utilisée pour rejouer les jeux : l'image jouée et la région cachée sont affichées avec la séquence d'indications ponctuées par des clics de l'enquêteur. La figure 8 illustre cette fonctionnalité.

Numbers : Un total de 3250 jeux constitue le résultat des deux campagnes de jeux. Parmi ces jeux, 3010 se sont terminés. Seuls 114 jeux ne se sont pas terminés avec avec la découverte de la région (96% des jeux sont gagnants avec un score, lié au temps, plus ou moins important). La durée moyenne d'un jeu se terminant par un succès est de 29 secondes. Cette durée passe à 33 secondes si on intègre les autres jeux.

Plus de 5000 clics sont collectés pour un total de 9063 indications. Il faut comprendre qu'un clic est associé à plus d'une indication au fil du jeu (les indices sont cumulatifs). Par exemple dans la figure 1, le clic victorieux est simultanément "on the right of the dog", "on the cat" et "on the cat's head". Ce clic apporte donc plus d'une indication.

Figure 8: *Les traces d'un jeu visualisées au travers de l'interface web : à gauche sont visibles l'image, la région cachée et les clics. A droite, les* tags *émis par l'enquêteur durant un jeu.*

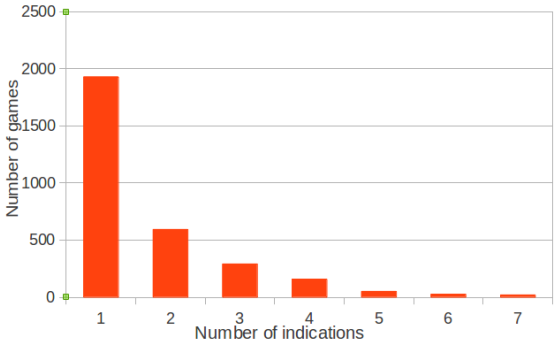


Figure 9: *Distribution du nombre d'indications par jeu.*

La figure 9 montre la distribution du nombre d'indications par jeu. Un grand nombre de jeux joués (pratiquement deux tiers) ont seulement une indication. Une stratégie coopérative efficace est évidemment de cacher la région au centre de l'objet principal ou dominant dans une image. Par exemple la figure 11 montre une situation où il est efficace de placer la région sur la tête d'un animal ou d'un humain présents dans l'image. Quand il y a plusieurs animaux (comme le chien et le chat) il faut statistiquement un peu plus d'une indication par jeu pour gagner.

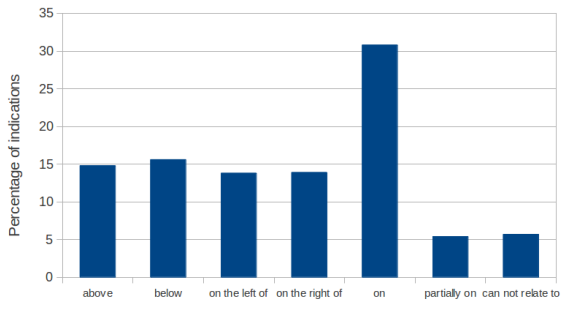


Figure 10: *Statistiques à propos des occurrences des relations spatiales.*

La figure 10 montre la distribution des relations spatiales indiquées par le maître à l'enquêteur. Il est intéressant de noter que les relations *on* sont de loin les plus utilisées. C'est une conséquence directe des faits expliqués ci-dessus : le maître a tendance à cacher la région sur une région saillante qui facilite le rôle de l'enquêteur et maximise le score. Les clics sur les objets les plus saillants sont donc majoritaires, ce qui est intéressant du point de vue de la compréhension d'images. Le second aspect intéressant est que les relations *above*, *below*, *on the left of*, et *on the right of* sont équitablement réparties. Ce qui signifie probablement que la position des objets les plus saillants vers lesquels le maître se penche sont plutôt uniformément répartis dans les images.

Figure 11: *Position des régions cachées dans l'image du chien et du chat.*

Exemples extraits des données. Pour bien montrer le potentiel des données collectées nous avons systématiquement fait joué les joueurs sur une image PASCAL (celle du chien et du chat). Cette image fait partie de celles qui sont assez adaptées au jeu, assez faciles "à gagner". Nous avons enregistré 99 jeux sur cette image particulière. La figure 11 montre les positions des 99 régions cachées. Il est notable que ces régions couvrent les deux deux objets saillants de l'image. Les densités sont plus fortes sur les têtes du chat et du chien. Nous observons logiquement que les deux *tags* les plus utilisés sont évidemment "cat" et "dog" avec respectivement 67 and 66 occurrences. Dans ce décompte nous n'ajoutons pas les mots "cat" et "dog" lorsqu'ils apparaissent dans des textes structurés comme "cat's head" ou "dog's leg".

La figure 12 montre tous les clics issus de tous les enquêteurs relativement au chien. Les clics en jaune sont "on the dog". Les clics en rouge sont "above", "below", "on the left of" ou "on the right of" du chien. On observe des erreurs. Par exemple, des clics théoriquement attendus sur le chien sont en dehors. Il y aussi aussi des points qui devraient être à l'arrière plan qui sont sur le chien. Un des sujets ouverts que nous soumettons à la communauté de recherche est l'élimination des ces erreurs soit via une approche robuste statistiquement soit avec des raisonnements plus sophistiqués intégrant, éventuellement, les propriétés de l'image. C'est une de nos pistes actuelles de recherche.

Figure 12: *Les clics des enquêteurs relativement au chien.*

7. Utilisations possibles du corpus

Le corpus peut être utilisé pour de nombreuses applications et pas uniquement dans le domaine de l'imagerie ou du multimédia. Voilà une liste non exhaustive des travaux qu'il serait pertinent d'envisager à partir des traces du jeu.

Analyses textuelles. Le traitement des informations textuelles collectées par notre jeu est un premier enjeu important. De très nombreuses inférences à propos des catégories d'images ou d'objets, à propos des instances multiples d'objets, à propos des parties d'objets seraient un plus évident dans une démarche de compréhension sémantique globale des images : S'agit-il d'une scène intérieure ou extérieure ? Y a-t-il plusieurs chats dans cette image ? La patte dont on parle serait-elle celle du chat ?

Détection d'objets. La contribution majeure des traces de notre jeu est à l'évidence l'association entre données textuelles et informations spatiales dans l'image. La manière d'opérer ces associations est originale car, selon notre connaissance de l'état de l'art, elle intègre des relations spatiales pour la première fois dans un jeu GWAP. Chaque clic contribue à contraindre la recherche d'un objet dans l'image. Il est naturellement possible d'augmenter la probabilité de détection d'un objet grâce à chaque triplet (*tag*, *spatial relation*, *click*) [CMC12] et aux effets cumulatifs des indications.

Une utilisation "asymptotique" du jeu (c'est-à-dire un cas où beaucoup de jeux sont joués sur chaque image) est possible en vue de l'établissement d'une vérité terrain pour la détection ou la segmentation d'objets.

Cependant et de manière plus réaliste, c'est dans un scénario semi-supervisé que le potentiel du corpus réside plutôt, selon nous. Il nous semble que l'intérêt des traces collectées est de semi-superviser des algorithmes de vision par ordinateur : éliminer des fausses détections d'un algorithme d'OpenCV, renforcer ou combiner des conclusions de détecteurs différents ? Tout cela grâce à une contribution peu coûteuse en nombre de jeux par image.

Elimination d'erreurs. Nous avons écrit plus haut que certaines erreurs (humaines) contaminaient le corpus et nous venons de souligner que les algorithmes de vision commettent aussi des erreurs. Ces imprécisions ou aberrations

qu'elles soient humaines ou algorithmiques s'avèrent de nature si différentes qu'il nous semble possible d'utiliser la vision pour corriger l'humain et réciproquement ! C'est un sujet intéressant vers lequel nous nous dirigeons. Une forme d'apprentissage actif semble aussi envisageable : un algorithme pourrait-il demander (via un jeu ou quelques jeux) une aide pour conforter sa conclusion ? des jeux pourraient-ils faire appel aux sorties d'algorithmes de vision pour détecter des traces aberrantes ?

Segmentation ou **Image Parsing**. La lecture (*parsing*) ou annotation sémantique complète d'une image est probablement l'application la plus intéressante de nos données. Ce corpus est un pas vers l'établissement d'une méthodologie de résolution semi-supervisée de ce problème, dès lors qu'on admet un humain dans la boucle d'interprétation. Les approches semi-supervisées sont ou bien vues comme des méthodologies supervisées avec peu d'exemples d'apprentissage ou bien vues comme des techniques non-supervisées avec des contraintes. Ces deux approches d'utilisation du corpus sont possibles.

8. Conclusion

En conclusion, nous livrons à la communauté un corpus de données visuelles et sémantiques qui possèdent un potentiel intéressant. Le corpus est accessible à l'URL suivante (http ://TBC). Ces données sont issues d'un jeu GWAP original que nous avons proposé il y a deux ans. Le jeu est en ligne à l'URL suivante (http ://TBC). Ce délai de deux ans a été nécessaire pour collecter un nombre significatif de traces qui sont aujourd'hui exploitables et qui pourraient alimenter des recherches au-delà de notre laboratoire. Nos propres travaux sur ces données sont partagés avec des collègues étrangers [CMC12, SCGiN*13]. Les auteurs remercient en particulier Ogé Marques (FAU, USA) et Xavi Giro-i-Nieto (UPC Barcelona) qui seront naturellement associés, dans le futur, à une version ou extension anglaise de cet article.

Références

[CMC12] CARLIER A., MARQUES O., CHARVILLAT V. : Ask'nseek : A new game for object detection and labeling. In *ECCV'12 Workshops*. 2012, pp. 249–258.

[EVGW*10] EVERINGHAM M., VAN GOOL L., WILLIAMS C. K., WINN J., ZISSERMAN A. : The pascal visual object classes (voc) challenge. *IJCV. Vol. 88*, Num. 2 (2010), 303–338.

[GRC*10] GULSHAN V., ROTHER C., CRIMINISI A., BLAKE A., ZISSERMAN A. : Geodesic star convexity for interactive image segmentation. In *CVPR'10* (2010), IEEE, pp. 3129–3136.

[LYS*11] LIU T., YUAN Z., SUN J., WANG J., ZHENG N., TANG X., SHUM H.-Y. : Learning to detect a salient object. *PAMI. Vol. 33*, Num. 2 (2011), 353–367.

[MFTM01] MARTIN D., FOWLKES C., TAL D., MALIK J. : A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV'01* (2001), vol. 2, IEEE, pp. 416–423.

[MO10] MCGUINNESS K., O'CONNOR N. E. : A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition. Vol. 43*, Num. 2 (2010), 434–444.

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A. : Grabcut : Interactive foreground extraction using iterated graph cuts. In *TOG* (2004), vol. 23, ACM, pp. 309–314.

[RTMF08] RUSSELL B. C., TORRALBA A., MURPHY K. P., FREEMAN W. T. : Labelme : a database and web-based tool for image annotation. *IJCV. Vol. 77*, Num. 1-3 (2008), 157–173.

[SCGiN*13] SALVADOR A., CARLIER A., GIRO-I NIETO X., MARQUES O., CHARVILLAT V. : Crowdsourced object segmentation with a game. CrowdMM '13, ACM, pp. 15–20.

[VAD04] VON AHN L., DABBISH L. : Labeling images with a computer game. In *CHI'04* (2004), ACM, pp. 319–326.

Identification du système d'acquisition scanner X à partir de l'analyse du bruit dans des images médicales

A. Kharboutly[1], W. Puech[1], G. Subsol[1] et D. Hoa[2]

[1]LIRMM
[2]IMAIOS

Résumé

L'imagerie médicale aide les médecins à améliorer et accélérer le processus de diagnostic. Il est donc fondamental de s'assurer que les images d'un patient n'ont pas été altérées ou interverties avec celles d'un autre. Pour cela, nous proposons une méthode pour identifier l'appareil scanner X à partir du bruit dans les images médicales. Nous avons construit un modèle de bruit de référence pour les images acquises par chaque système scanner X. Nous avons ensuite corrélé les images 3D obtenues avec chaque modèle de bruit de référence pour identifier l'appareil scanner X correspondant. Nous avons utilisé une approche de filtre de Wiener basé sur des ondelettes pour extraire le bruit. Des résultats expérimentaux préliminaires ont été obtenus sur 8 images 3D de 100 coupes de scanner X différents et nous avons pu globalement identifier chaque scanner X.

Medical image processing is used to help the doctors to improve and accelerate the diagnostic process. Consequently, it is essential to ensure that the images of a patient were not altered or swapped. Therefore, we propose a method for CT-Scan identification based on the sensor noise pattern. We extracted the reference noise pattern for each CT-Scan from its 3D image, and then we correlated the tested 3D images with each reference noise pattern in order to identify the corresponding CT-Scan. We used a wavelet-based Wiener filter approach to extract the noise. Experimental results were applied on 8 3D image of 100 slices that were around 800 slices from 3 CT-Scans. Generally, we were able to identify each CT-Scan separately.

Mots clé : criminalistique numérique, imagerie médicale, authentification, identification des dispositifs, analyse de bruit.

1. Introduction

L'imagerie médicale est devenue de nos jours un enjeu essentiel dans le monde de médecine, elle offre des techniques qui peuvent être utilisées pour regarder à l'intérieur de l'organisme de manière non intrusive. Le traitement d'image médicale est devenu une technique commune dans le domaine du traitement d'images.

La tomographie [BSJB11] appelée aussi scanner X, intègre une série des vues radio-graphiques qui sont prises sous des angles différents pour créer des images 3D des os et des tissus mous de l'intérieur de l'organisme. Elles peuvent être utilisées pour visualiser toutes les parties du corps et elles sont largement utilisées puisqu'elles fournissent beaucoup d'informations sur les caractéristiques physiques et les pathologies du patient. Il est donc fondamental de s'assurer que ces données ne sont pas altérées ou interverties avec celles d'un autre patient.

Les informations sur l'acquisition et l'identification du dispositif, sont généralement stockées dans des fichiers DICOM [Toe12]. Le fichier DICOM peut être décomposé en deux parties, les meta-données et l'image brute. Les meta-données sont facilement lisibles, elle contiennent toutes les informations sur le dispositif d'acquisition, les paramètres d'imagerie et plus généralement la procédure suivie. Mais si ces méta-données sont dissociées de l'image volontairement ou involontairement (par un changement de format par exemple), il est important de pouvoir retrouver un maximum d'informations à partir de l'image elle-même. En particulier l'identification précise de système d'acquisition permettra de remonter au centre de radiographie et de pouvoir y retrouver dans les archives les paramètres de l'image.

En leur absence ou si les meta-données sont non authentifiées, nous pouvons identifier le scanner X à partir des images brutes. C'est justement l'objectif de la criminilastique d'images, un domaine de la recherche important [RTD11], qui a pour but de valider l'authenticité des images en récupérant des informations sur leur histoire, en présence du dispositif non authentifié ou de la modification de contenu. En termes de criminalistique d'images, deux problèmes sont abordés : le traçage de contrefaçons et l'identification des dispositifs d'acquisition. Dans le cas de la contre-

façon, beaucoup de travaux sont produits en photographie numérique générale [SM13]. Dans le cas des images médicales et plus spécifiquement des scanner X, très peu de recherches ont été effectuées. Dans [HCS*12], les auteurs présentent un premier travail en criminalistique numérique aveugle dans le domaine de l'imagerie médicale. Ils ont proposé une méthode permettant de détecter si une image a été modifiée ou non grâce à des opérateurs de traitement d'image. Pour l'identification du dispositif, aucune méthode n'a été proposée pour l'imagerie scanner X, bien qu'il y ait quelques travaux sur l'analyse des caractéristiques d'image médicales selon les paramètres d'acquisition et le dispositif [SCS12]. Dans [LFG06], les auteurs ont proposé une méthode pour l'identification d'appareil photo numérique du bruit de modèle de capteur, ils ont utilisé un algorithme de débruitage en ondelettes permettant de séparer la composante de bruit. Puis, ils ont généré un des motifs de bruit de référence pour l'appareil photo numérique et, enfin, ils ont utilisé la corrélation pour mettre en correspondance l'image avec un appareil photo. Cette méthode est appliquée aux images photographiques mais elle pourrait être généralisé aux coupes des images 3D obtenues pas scanner X.

Dans cet article, nous proposons une première analyse du problème d'identification du système d'acquisition scanner X à partir des coupes 2D. Le reste de ce papier est organisé comme suit. Dans la Section 2, nous présentons un algorithme de débruitage, nous avons construit un modèle de bruit de référence pour chaque dispositif, nous avons identifié le scanner X grâce à la corrélation entre les coupes testées et le modèle de bruit de référence de chaque dispositif. Dans la Section 3, nous exposons nos résultats expérimentaux et nous les commentons et finalement, nous concluons notre travail avec des perspectives dans la Section 4.

2. Méthode d'Identification de scanner X

La méthode proposée est basée sur la méthode présentée dans [LFG06]. comme plusieurs machines (scanner XS) ont été utilisées, nous avons crée une référence de bruit pour chaque appareil comme présenté dans la figure 1. La corrélation entre la composante de bruit de l'image et la référence de bruit de chaque machine a été étudiée pour que nous puissions identifier une image acquise pour chaque machine. Cette image est considérée comme acquise par une certaine machine quand il y a une valeur de corrélation élevée avec sa référence de bruit comme illustré dans la figure 1. Dans cette partie nous présentons l'algorithme qui a été utilisé pour isoler le bruit, puis la référence de bruit qui a été créé et finalement, comment la décision de l'identification de la machine est faite.

2.1. Algorithme de débruitage

Nous avons appliqué un filtre en utilisant une transformation en ondelettes dans le domaine des fréquences et basé sur le travaux proposés dans [MKR99]. Essentiellement, il est composé de deux parties, l'estimation de la variance locale des composantes d'ondelettes et le débruitage de ces composantes en utilisant un filtre de Wiener [JM04] comme suit :

- Calculer quatre niveaux de décomposition en ondelettes. Dans chaque niveau, marquer les trois sous-bandes de haute fréquence qui sont horizontale, verticale et la diagonale. Pour quatre niveaux de décomposition en ondelettes avec trois sous-bandes dans chaque niveau, nous avons donc 12 sous-bandes pour chaque image traitée
- Pour chaque sous-bande d'ondelettes, nous estimons la variance locale en utilisant un voisinae de (3x3) à (9x9) en fonction du niveau. :

$$\hat{\sigma}^2_W(i,j) = max\left(0, \frac{1}{W^2}\sum_{(i,j)\in W*W}\left(X^2(i,j) - \sigma_0^2\right)\right), \tag{1}$$

où $W \in \{3,5,7,9\}$ se réfère à la taille du voisinage, X est la sous-bande d'ondelette et σ_0 est une valeur constante donnée.
Parmi les valeurs précédentes correspondant aux 4 niveaux de voisinage, nous choisissons la valeur minimale :

$$\hat{\sigma}^2(i,j) = min\left(\sigma_3^2(i,j), \sigma_5^2(i,j), \sigma_7^2(i,j), \sigma_9^2(i,j)\right). \tag{2}$$

- Débruitage des sous-bandes d'ondelettes utilisant le filtre de Wiener :

$$X_{den}(i,j) = X(i,j)\frac{\hat{\sigma}^2(i,j)}{\hat{\sigma}^2(i,j) + \sigma_0^2}, \tag{3}$$

où X est la sous-bande d'ondelettes.
- Appliquer la transformation inverse ondelettes sur les sous-bandes d'ondelettes débruitées pour obtenir le composant débruité $F(s)$ de l'image originale s.

2.2. Modèle de référence de scanner X

Pour chaque appareil scanner X, nous avons extrait des images 2D correspondant à des coupes de l'image 3D. Pour chaque groupe d'images, nous avons appliqué un filtre de débruitage pour extraire le signal de base, puis nous avons soustrait l'original de chaque coupe comme représenté dans la figure 2 :

$$n^{(i)} = s^{(i)} - F(s^{(i)}), \tag{4}$$

où n représente la composante de base du bruit, s la coupe numéro i et $F()$ est la fonction de débruitage de la coupe numéro i.

Nous avons ensuite moyenné les bruits en une seule image 2D. C'est ce que l'on appelle la ligne de base du bruit où encore la signature de l'appareil. En répétant le processus avec chaque groupe d'images, on obtient la référence complète de l'appareil :

$$RPN = \frac{1}{N}\sum_{i=1}^{N} n^{(i)}, \tag{5}$$

où RPN est le bruit de référence, N le nombre de coupes dont le bruit est extrait et n la composante du bruit.
Comme indiqué dans [GBdG04], nous supposerons que le bruit est additif gaussien.

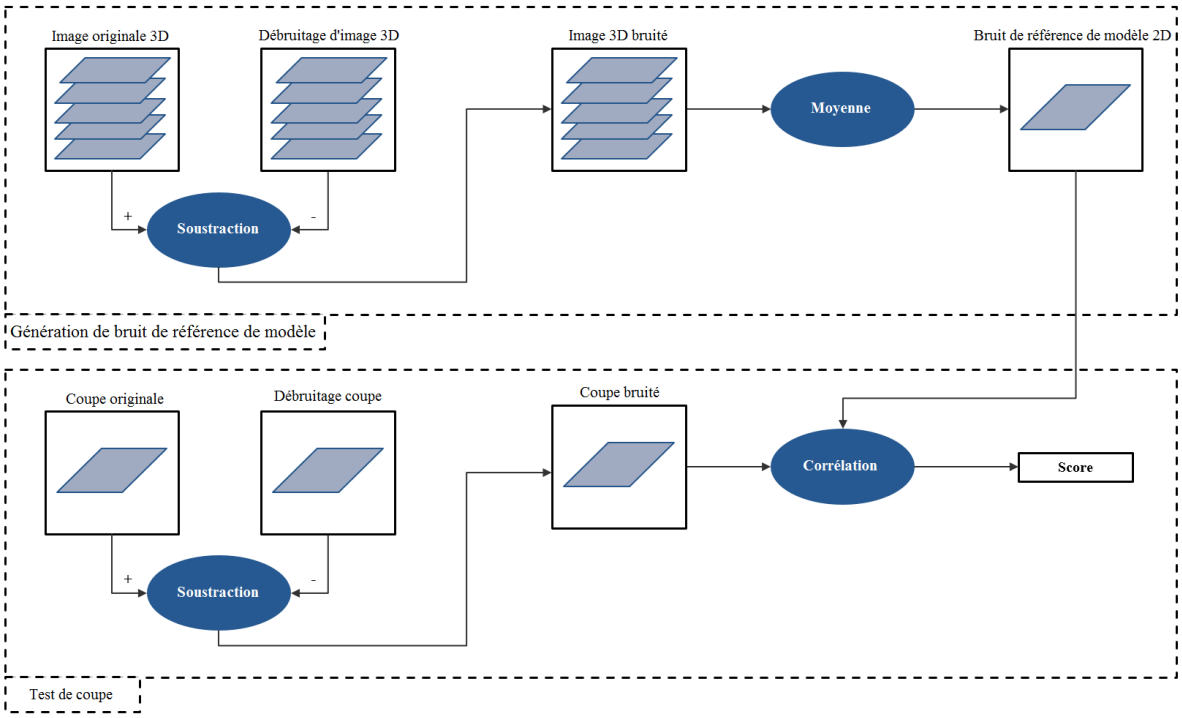


Figure 1: *Vue d'ensemble de la méthode.*

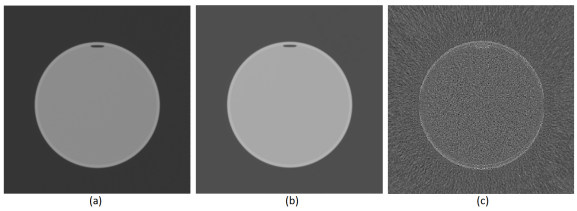


Figure 2: *a) Exemple d'une coupe initiale de Siemens 1, b) composante débruitée, c) la composante de bruit.*

2.3. Décision par corrélation

Nous souhaitons identifier l'appareil à partir duquel chaque image de chaque groupe a été acquise. Pour cela, nous extrayons le bruit de référence de chaque appareil et les composantes du bruit de chaque image. Ces images sont reliées à un des appareils lorsque leur coefficient de corrélation avec le bruit de référence de l'appareil est maximal :

$$corr(n^{(i)}, RPN) = \frac{(n^{(i)} - \overline{n^{(i)}}).(RPN - \overline{RPN})}{\|n^{(i)} - \overline{n^{(i)}}\| \|RPN - \overline{RPN}\|}. \quad (6)$$

3. Résultats Expérimentaux

Pour tester cette méthode, nous avons appliqué le processus sur 8 images 3D issues de 3 scanner X différents de 2 constructeurs : Siemens (siemens 1 et siemens 2) et General Electric Medical Systems (GE). Ces images sont codées en 16 bits et acquises en utilisant les mêmes paramètres (énergie du faisceau=120 kV ; pas de 1 et épaisseur de coupe de 3 mm). Chaque image 3D est composée de 100 coupes de 512*512 pixels. Trois images 3D de fantôme de crâne adulte ont été acquises avec Siemens1, trois images 3D de fantôme de crâne adulte de Siemens 1 et 2 ; ainsi que 2 images 3D de crâne de General Electrics comme résumé dans le Tableau 1.

	Siemens 1	Siemens 2	GE
Content	fantôme	fantôme	crâne
Nb d'images 3D	3	3	2
Nb de coupes	300	300	200
Nb de coupes de RPN	120	120	120
Nb de coupes testées	180	180	80
Taille (pixels)	512x512	512x512	512x512
Bits par pixel	16	16	16
Épaisseur des coupes	3 mm	3 mm	3 mm
Taille du pixel	1 mm	1 mm	1 mm

Table 1: *Caractéristiques des images expérimentales.*

En fonction des différentes méthodes d'extraction RPN citées ci-dessus, nous avons extrait un bruit de référence pour chaque appareil. La figure 3 illustre les 3 références pour chaque appareil. On peut remarquer les composantes de bruit en plus de quelques effets de bords, car il y a des contours dans les images originales. Les bords de certaines structures restent apparents dans l'image quand on moyenne les coupes et on peut remarquer leurs contours dans la figure 3.

Le reste des coupes de chaque appareil est conservé pour tester l'identification scanner X. Ces coupes sont testées

avec le motif de référence de bruit de chaque appareil. 120 coupes GE, 180 de Siemens 1 et 180 de Siemens 2. Pour confirmer les résultats, nous avons répété l'expérience 5 fois avec des coupes aléatoires et le résultat obtenu est le même.

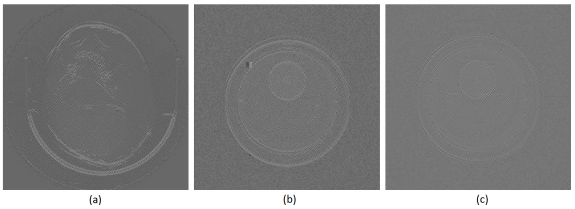


Figure 3: *Exemple du bruit de motif de référence à partir de : a) General Electric, b) premier dispositif de Siemens, c) deuxième dispositif de Siemens.*

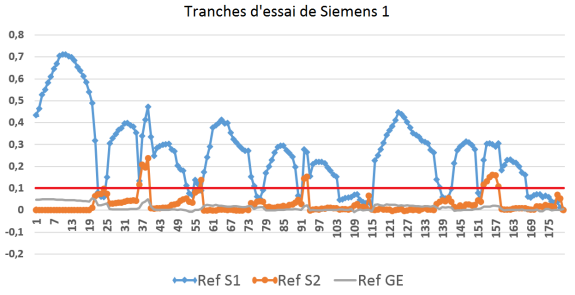


Figure 4: *Les corrélations entre les coupes testées de Siemens 1 et le motif de bruit de référence en ce qui concerne chaque appareil.*

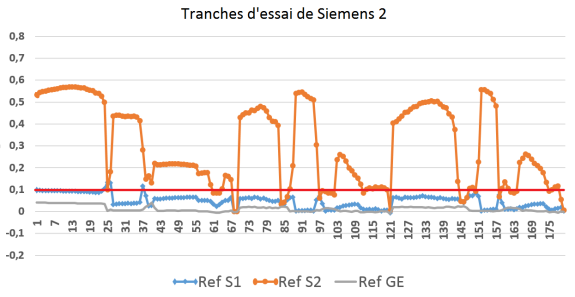


Figure 5: *Les corrélations entre les coupes testées de Siemens 2 et le motif de bruit de référence en ce qui concerne chaque appareil.*

Les courbes dans la figure 4, figure 5 et figure 6 permettent de remarquer que la corrélation entre les coupes testées et les bruits de référence de l'appareil concerné est la plus élevée. La figure 4 illustre la corrélation entre les 3 bruits de référence (S1, S2 et GE) et les 180 coupes de S1. L'axe vertical correspond à la corrélation et l'axe horizontale au numéro de coupe. Concernant l'axe vertical, on peut remarquer la corrélation entre la référence de Siemens 1 et les coupes testées

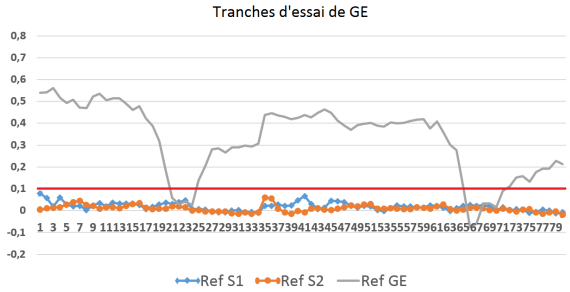


Figure 6: *Les corrélations entre les coupes testées de General Electric et le motif de bruit de référence en ce qui concerne chaque appareil.*

de Siemens 1, la quasi-totalité de ces valeurs de corrélation sont supérieures à 0,1, tandis que toutes les autres valeurs de corrélation sont inférieures, qui se réfère à la relation entre ce dispositif et de l'image test. Ainsi, on peut considérer la valeur 0,1 comme seuil qui classe les images étant acquises par le dispositif de cette référence, sauf quelques images concernant son contenu, comme nous pouvons le constater dans le Tableau 2. Nous avons aussi les même remarques pour les figure 5 et figure 6.

	Siemens 1	Siemens 2	GE
Siemens 1	95.5 %	3.0 %	5.0 %
Siemens 2	4.0 %	97.0 %	0
GE	0.5 %	0	95.0 %

Table 2: *Précision d'identification*

Le tableau 2 montre le taux de classification, lorsque nous avons corrélé 180 coupes de Siemens 1, 180 coupes de Siemens 2 et 80 coupes de General Electric avec le bruit de référence de chaque appareil séparément.

4. Conclusion et perspectives

Dans cet article, nous avons proposé un algorithme pour la criminalistique d'images médicales. Pour les études à venir, travailler avec plus d'images et de dispositifs pour valider notre approche, essayer de généraliser notre travail en 3D, étudier l'influence des paramètres d'acquisitions et de reconstruction sur le bruit, étudier l'influence du contenu de l'image sur le calcul du bruit en particulier pour supprimer les artefacts des contours, analyser le processus de reconstruction pour mieux modéliser le bruit, étudier les caractéristique des coupes qui donnent faible corrélation et étudier l'influence de la compression sur l'image de référence.

Références

[BSJB11] BUSHBERG J. T., SEIBERT J. A., JR. E. M. L., BOONE J. M. : *The Essential Physics of Medical Imaging, Third Edition*, third, north american edition ed. LWW, 12 2011.

[GBdG04] GRAVEL P., BEAUDOIN G., DE GUISE J. A. : A method for modeling noise in medical images. *IEEE Trans. Med. Imaging. Vol. 23*, Num. 10 (2004), 1221–1232.

[HCS*12] HUANG H., COATRIEUX G., SHU H., LUO L., ROUX C. : Blind integrity verification of medical images. *IEEE Transactions on Information Technology in Biomedicine. Vol. 16*, Num. 6 (2012), 1122–1126.

[JM04] JACOB N., MARTIN A. : Image denoising in the wavelet domain using Wiener filtering. [Online], Project Report, Available : http ://homepages.cae.wisc.edu/ ece533/project/f04/.

[LFG06] LUKAS J., FRIDRICH J., GOLJAN M. : Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security. Vol. 1*, Num. 2 (2006), 205–214.

[MKR99] MIHÇAK M., KOZINTSEV I., RAMCHANDRAN K. : Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on* (Mar 1999), vol. 6, pp. 3253–3256.

[RTD11] REDI J., TAKTAK W., DUGELAY J. L. : Digital image forensics : a booklet for beginners. *Multimedia Tools and Applications. Vol. 51*, Num. 1 (2011), 133–162.

[SCS12] SOLOMON J. B., CHRISTIANSON O., SAMEI E. : Quantitative comparison of noise texture across CT scanners from different manufacturers. *Medical physics. Vol. 39*, Num. 10 (October 2012), 6048–55.

[SM13] SENCAR H. T., MEMON N. : *Digital Image Forensics : There is More to a Picture Than Meets the Eye*. Springer, 2013.

[Toe12] TOENNIES K. D. : *Guide to Medical Image Analysis - Methods and Algorithms*. Advances in Computer Vision and Pattern Recognition. Springer, 2012.

Nouvelle méthode d'évaluation de robustesse des algorithmes de tatouage vidéo: Jeu d'attaque

Asma Kerbiche[1], Saoussen Ben Jabra[1], Ezzeddine Zagrouba [1], Axel Carlier[2,3] et Vincent Charvillat[3]

[1]Université Tunis El Manar
Lab. RIADI - Equipe SIIVA

[2]Université de Toulouse
Lab. IRIT - Equipe Vortex

[3]Université de Singapore

Résumé
L'évaluation d'une technique de tatouage a été toujours une étape critique et importante. En effet, l'évolution progressive des outils de traitement et de communication de vidéos a fait naître de nombreuses et différentes techniques de tatouage dont l'efficacité varie d'un algorithme à un autre. Cette efficacité est toujours évaluée en se basant sur plusieurs contraintes dont les plus importantes sont l'invisibilité et la robustesse face aux attaques. Cette dernière est souvent évaluée en testant des attaques classiques et simples telles que la compression, la rotation, la translation et l'ajout de bruit. Des techniques simples de tatouage peuvent résister à des attaques simples sans pour autant être robustes aux attaques observées dans le monde réel comme le "Camcording" d'un contenu vidéo. En situation réelle, un utilisateur mal intentionné (ou un pirate) va filmer illégalement un contenu projeté, recadrer l'image, transcoder le contenu obtenu. La question que nous nous posons est l'évaluation de méthodes de tatouage dans ce type de contexte. Dans le cadre de l'élaboration d'un nouveau protocole d'évaluation de techniques de tatouage vidéo, nous proposons, dans ce papier, un jeu d'attaques de vidéos tatouées mis à disposition d'un ensemble d'utilisateurs qui vont interagir afin de détruire la marque qui a été insérée. Ceci sera réalisé en leur fournissant une liste d'attaques qu'ils peuvent librement appliquer et combiner à ces vidéos tatouées. Cette liste va comprendre les attaques les plus importantes et réelles que peut subir une vidéo telles que le camcording, la déformation, l'ajout de couleur et la compression. Ce jeu nous a permis non seulement d'évaluer n'importe quel algorithme de tatouage vidéo, mais surtout, d'identifier, à partir de l'étude de choix des utilisateurs les attaques les plus importantes pour eux.

Résumé
Watermarking techniques evaluation presents a critical and very important step. In fact, with the progressive evolution of video processing and communication, many watermarking approaches are proposed and their efficiency varies from an approach to another one. This efficiency is generally evaluated based on several constraints where the most important are invisibility and robustness in front of attacks. This last one is often estimated by testing classical and simple attacks such as rotation, translation and noise add. Nevertheless, a simple watermarking technique can easily resist to this type of attacks. So, it is necessary to study the impact of other attacks which are always applied in reality such as camcording attack. The integration of these important attacks will allowed a more efficient comparison between the various proposed techniques. In this paper, a new attacking game is proposed. In this game, different marked videos will be at the disposal of a set of users who can interact to destroy the inserted mark. Indeed, a list of different attacks will be given to users who can apply a combination of several of them to the marked videos. This list will contain the most important and dangerous attacks such as camcording, color add compression. This game allowed us to evaluate our video watermarking technique. In more, based on users' choices, we can identify the most important attacks for them.

Mots clé : tatouage, vidéo, crowdsourcing, camcording, attaques, jeu...

Mots clé : Watermarking, video, crowdsourcing, camcording, attacks, game...

1. Introduction

Avec l'avènement de l'ère numérique, l'information est devenue volatile et facilement interceptée et produite. Le

développement des réseaux à haut débit, notamment Internet, et l'évolution des nouvelles normes de décompression ont facilité la transmission et le partage de l'information. En conséquence, il est devenu très aisé de gérer les données volumineuses en termes de stockage et de traitement puisque il est devenu possible maintenant de faire des vidéo conférences en temps réel, d'envoyer par e-mail des fichiers de taille importante, de regarder des films depuis un serveur distant, etc. Les documents numériques sont donc soumis à plusieurs problèmes tels que le piratage et le non-contrôle de la copie qui peuvent provoquer une répercussion économique non négligeable. Il est devenu donc nécessaire pour les créateurs de contenus numériques de rechercher des solutions afin de lutter contrer ces problèmes. Le tatouage, appelé aussi "Watermarking" en anglais, s'avère une technique de sécurisation des données qui permet de remédier à ces problèmes. En effet, il permet d'insérer dans un document numérique une signature non perceptible puis de tenter à la récupérer après d'éventuelles attaques subies par le document tatoué. Nous nous sommes intéressés dans notre travail au tatouage des flux vidéo. Ce type de tatouage est un domaine assez récent qui a rencontré un intérêt croissant au sein de la communauté scientifique. En effet, de multiples méthodes de tatouage vidéo sont apparues et chacune d'elles possède ses avantages et ses inconvénients mais aucune ne parvient encore à s'imposer. En effet, un bon algorithme de tatouage doit obéir à deux principales contraintes : l'invisibilité de la signature insérée et sa robustesse contre les attaques. Cependant, malgré l'évolution des algorithmes de tatouage et des techniques de piratage, la robustesse de chaque algorithme de tatouage est toujours évaluée en testant les mêmes attaques 'usuelles' classiques. En effet, de nos jours on ne retrouve plus d'algorithme de tatouage qui n'est pas robuste face à la rotation ou l'ajout de bruit, cela fait partie des bases d'un bon algorithme de tatouage... Pour autant la résistance à ces attaques "artificielles" n'est pas équivalente à la robustesse face aux pratiques malveillantes observées dans le monde réel. Nous nous intéressons une attaque très répandue appelée « camcording » qui consiste en l'enregistrement des films dans les salles de cinéma à l'aide d'un Smartphone ou d'une caméra et qui peut être suivie de transformations colorimétriques, d'une compression ou d'une déformation ... Ces attaques doivent alors être prises en considération dans le processus d'évaluation des techniques de tatouage vidéo. Afin d'élaborer cette évaluation, nous avons eu recours au crowdsourcing qui permet d'utiliser l'intélligence collective de beaucoup d'utilisateurs. Nous avons conçu un jeu d'attaques o*ù* nous avons fait interagir des utilisateurs qui vont essayer de détruire le tatouage inséré dans la vidéo et ceci en appliquant plusieurs combinaisons d'attaques disponibles, tout en gardant une bonne visibilité de la vidéo attaquée. Cette interaction va nous permettre de savoir quelles sont les attaques les plus intéressantes pour les utilisateurs. La suite de ce papier sera organisée comme suit : dans la section suivante, un état de l'art concernant le tatouage vidéo, les protocoles d'évaluation existants ainsi que la technique du crowdsourcing sera présenté. La deuxième partie sera consacrée à la description du jeu d'attaques proposé tout en exposant la liste d'attaques choisies pour l'évaluation. Les résultats obtenus feront l'objet de la troisième section et nous finirons par une conclusion et certaines perspectives.

2. Evaluation des techniques de tatouage vidéo

Tout algorithme de tatouage doit obéir à certaines contraintes dont les plus importantes à prendre en compte pour l'évaluation sont l'invisibilité, la capacité et la robustesse cette derniére étant la plus critique. En effet, la marque doit être capable de résister à plusieurs types d'attaques. On appelle "attaque" une transformation que l'on va faire subir à l'image et qui va plus ou moins l'endommager mais qui risque d'être fatale au marquage [A.P99]. La robustesse d'un marquage dépend de sa capacité à résister à une attaque. En fait, cette attaque peut avoir deux buts : le premier est d'accentuer ou de masquer certaines caractéristiques de l'image. Le deuxième est de rechercher la marque et la lire, la détruire ou la changer de façon à rendre impossible sa détection [FJB99]. La détection doit alors être possible quelque soit l'attaque appliquée. Pour la vidéo, les attaques les plus utilisées pour les tests de robustesse sont des attaques simples que nous jugeons pas assez réalistes et classiques qui peuvent être aussi classées selon l'intention de leurs utilisations. Dans ce cas, nous distinguons deux grandes familles : les attaques innocentes comme la compression, les transformations géométriques (Translation, rotation, changement d'échelle...), le filtrage ou l'ajout de bruit et les attaques malveillantes telle que la collusion qui consiste essentiellement à moyenner les images successives d'une vidéo dans le but d'éliminer la marque sans nuire à la qualité de la séquence.

Plusieurs protocoles ont été proposés pour les algorithmes de tatouage vidéo, cependant, ils utilisent généralement les mêmes attaques classiques pour l'évaluation de la robustesse. Parmi ces protocoles, nous pouvons citer le projet Européen Certimark [Rol] qui a été lancé en mai 2001 sous la direction de C. Rollin. Ce protocole travaille principalement sur la réalisation de tests génériques pour l'évaluation des méthodes de tatouage d'images et vidéo. Les attaques testées pour les vidéos sont : la compression, la conversion numérique en analogique et analogique en numérique (D/A et A/D), application des formats de stockage de la vidéo avec perte, ajout de logos ou sous-titrage, transformations géométriques (rotation, translation, Cropping, changement d'échelle...), ajout d'autres marques, bruit et collusion. Plusieurs autres protocoles d'évaluation ont été proposés pour les algorithmes de tatouage d'image tel que le projet Stirmark benchmark [Pet00] qui a été lancé en 1998 proposé par Fabien A. P. Petitcolas, le projet « BOWS : Break Our Watermarking System » [BF] qui a été lancé en 2007 qui évaluent la performance d'un algorithme de tatouage en se basant sur plusieurs contraintes (capacité, invisibilité, rapidité, robustesse...) et qui teste la vulnérabilité de ces algorithmes face aux attaques standards certes importantes mais qui ne présentent plus de risque pour les algorithmes de tatouage actuels.

3. La technique du Crowdsourcing

Avec le développement et le grand progrès qu'ont vécus les technologies Web 2.0, de nombreux systèmes so-

ciotechniques ont attiré l'attention des praticiens et des universitaires. Le Crowdsourcing est un nouveau phénomène émergent du Web 2.0 qui est devenu un mécanisme d'approvisionnement reconnu pour résoudre les problèmes des organisations et sociétés par l'externalisation de ces problèmes à une foule. Pour cela, le domaine du Crowdsourcing est devenu un secteur de recherche très dynamique et est en pleine croissance au fil de ces années. Le terme Crowdsourcing a été inventé par Howe, dans un article de Wired Magazine en juin 2006 [J.06] : "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers". Le Crowdsourcing est basé sur un simple mais puissant concept : Presque tout le monde a un potentiel pour diffuser des informations précieuses [GF11]. En effet, il s'agit de mobiliser les compétences et l'expertise qui sont réparties dans la foule [C.08]. Le Crowdsourcing n'est pas seulement utilisé pour des fins commerciales. En effet, de nombreuses organisations à but non lucratif l'ont adapté comme un modèle efficace pour la résolution de problèmes [C.10] [J.08]et ca a également attiré l'attention de la communauté universitaire. En effet, plusieurs études récentes sont basées sur la technique du Crowdsourcing. Xie & al. [XLGyM05] ont proposé une nouvelle méthode pour détecter les cartes d'intérêt des utilisateur en se basant sur le crowdsourcing cette méthode à montré une meilleure efficacité que celles basées sur l'analyse d'image, en représentant l'intérêt réel des utilisateurs. Carlier & al. [CCOM10] aussi ont proposé une méthode basée sur le Crowdsourcing qui déduit les régions d'intérêts d'une vidéo en analysant le comportement de visualisation implicite d'un grand nombre d'utilisateurs en utilisant le zoomin. Notre approche méthodologique, dans cet article, consiste dans la même voie à collecter des traces d'utilisateurs à qui nous demandons d'attaquer des vidéos tatouées. Nous motivons ces utilisateurs en leur lançant une forme de défi ludique : "Trouver la meillure attaques qui supprime la marque sans trop altérer la vidéo.

4. Jeu d'attaques proposé

Nous avons ainsi construit un "jeu" d'attaques que nous livrons à une foule d'utilisateurs/joueurs. La conception a été guidée par une attaque réelle appelée Camcording. Le camcording revient à filmer la vidéo affichée ou projetée à l'aide d'un Smartphone ou une caméra et puis à la diffuser après lui avoir appliqué certaines transformations afin de détruire la marque qui a été insérée. En effet, la copie illicite des films est une grande préoccupation de l'industrie cinématographique et le développement technologique au cours de ces dernières années a fait du piratage une menace encore plus grande. L'unité de recherche et développement de la société Kodak s'est rendue compte de l'importance de cette attaque et a conçu un algorithme de tatouage qui a montré ses preuves face au camcording et qui permet aussi d'identifier la salle de cinéma où la projection a eu lieu ainsi que l'heure et la date de diffusion [CMR01]. Philipp Schaber & al. [SKWE14], ont conçu un outil qui simule une ré-acquisition d'un contenu avec un caméscope pour soutenir les recherche comme la notre. Partant de ces travaux, nous avons choisi de concevoir un jeu d'attaques en se basant sur la technique du crowdsourcing. En effet, il s'agit de mettre à la disposition des utilisateurs une interface qui leurs permet d'appliquer à des vidéos tatouées un ensemble d'attaques que nous avons jugées les plus importantes et dangereuses. L'interface proposée est illustrée dans la figure 1, elle comporte trois parties principales : la première contient un aperçu de la vidéo originale, la deuxième contient la liste des attaques choisies et la dernière montre un aperçu de la vidéo tatouée. Chaque utilisateur a droit à trois essais au cours desquels il va tenter de détruire la signature. Il pourra appliquer les combinaisons qu'il désire à condition qu'il ne détériore pas trop la qualité visuelle de l'image attaquée.

En se basant sur des enquêtes faites avec des spécialistes de cinéma, nous avons choisi d'intégrer dans l'interface proposée les attaques suivantes : le camcording, la compression, la déformation, le cropping et la modification de couleur.

4.1. Le camcording

La première étape que peut appliquer un utilisateur en accédant au jeu est de choisir soit une vidéo originale soit celle camcordée. Pour ce faire, nous avons choisi de camcorder la vidéo à partir de 4 prises de vue. Au début nous avons positionné le caméscope en face de l'écran, puis à droite, puis à gauche et pour finir en bas comme le montre la figure 2. Par conséquent, chaque utilisateur peut choisir de travailler sur la vidéo originale ou une de ces quatre versions camcordées.

4.2. La compression

La compression MPEG4 est la norme de compression la plus populaire et la plus utilisée de nos jours. Ce format de compression est parfaitement adapté à la haute définition qu'il diffuse sans prendre trop de place sur le vecteur utilisé (satellite, câble, TNT).Vu l'importance de ce système de compression, tout bon algorithme de marquage doit pouvoir lui résister au moins dans les faibles taux de compression. C'est pour cette raison, que la deuxième étape dans notre jeu d'attaques consiste à permettre aux utilisateurs de choisir de travailler avec la version compressée ou non de la vidéo tatouée. Ils auront alors le choix entre 3 débits de compression (1000 kbit/s, 500 kbit/s et 200 kbit/s). La figure 3 présente la vidéo compressée pour chaque débit.

4.3. La déformation

Après le choix de la compression, les utilisateurs peuvent appliquer, dans le cas des vidéos camcordées, des modifications pour recadrer ces vidéos (figure 4) et ceci en sélectionnant les coordonnées de la région qu'ils veulent rectifier. Une prise de vue latérale en Camcording nécessite en effet un recalage homographique de l'image.

4.4. Le cropping

L'attaque du cropping ou rognage d'une séquence vidéo consiste à en extraire un morceau. Elle permet de couper horizontalement ou verticalement les images de la vidéo. Cette

Figure 1: *Interface du jeu d'attaques proposé.*

attaque peut détruire totalement la marque. Dans le jeu d'attaques proposé, les utilisateurs peuvent sélectionner la région de la vidéo qu'ils désirent conserver à condition de garder une bonne visibilité de la vidéo (figure 5).

4.5. L'ajout de couleur

La dernière attaque disponible dans le jeu proposé est l'ajout de couleur qui revient à modifier les couleurs de la vidéo en rajoutant soit du rouge, du vert, ou du bleu aux couleurs initiales des images de la vidéo (figure 6).

5. Résultats expérimentaux

Afin d'étudier et d'analyser les choix des différents utilisateurs participant au jeu proposé ainsi que le comportement du tatouage à tester, nous avons choisi d'évaluer l'algorithme de tatouage basé sur l'insertion multi-fréquentielle dans les régions d'intérêts que nous avons proposé dans nos travaux précédents [KJZ12] et la méthode proposé par Chan & al [CL03] qui propose un algorithme de tatouage vidéo basé sur la transformée en ondelette. Ces deux algorithme on présenté une bonne robustesse en les évaluant face aux attaques usuelles de tatouage vidéo comme la rotation, l'ajout de bruit, al suppression d'images et la compression... Ces al-

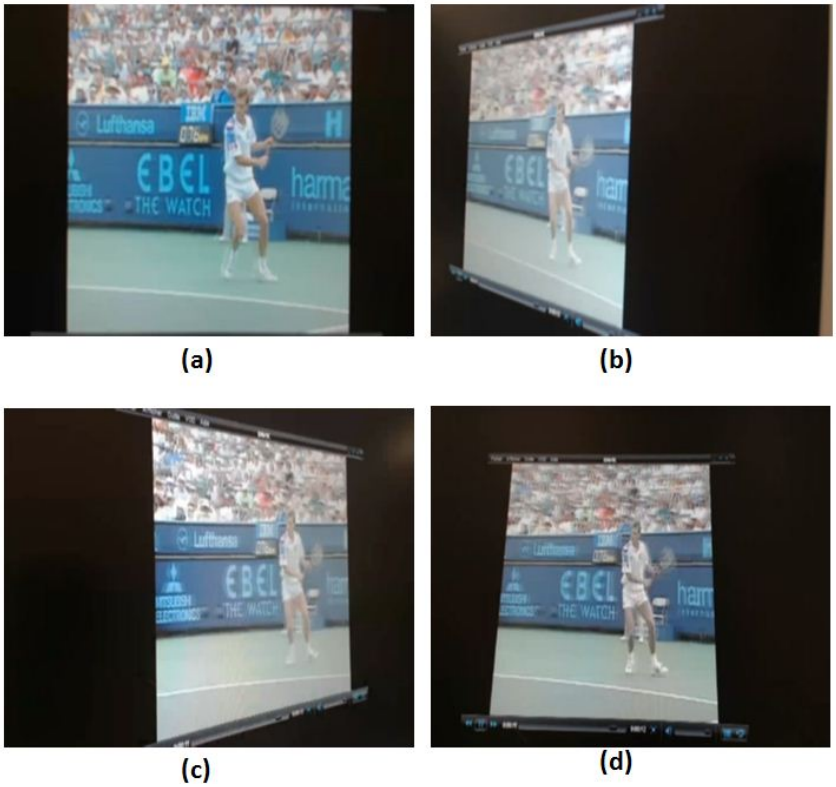


Figure 2: *Prises de vue : (a) caméra en face, (b) caméra à droite, (c) caméra à gauche, (d) caméra en bas.*

gorithmes ont été alors appliqués sur la séquence vidéo couleur Stefan composée de 300 images. La vidéo résultante a été mise à disposition de 20 utilisateurs qui vont essayer de détruire sa marque. Ces utilisateurs auront droit à trois essai pour chaque essai ils auront la possibilité d'appliquer autant d'attaques qu'ils veulent mais à condition de conserver une bonne qualité visuelle de la vidéo attaquée. En effet, après chaque choix d'attaque par l'utilisateur, un critère d'invisibilité est calculé et comparé à un seuil afin de valider ou non l'ensemble d'attaques choisies par l'utilisateur. Pour ce faire, nous avons utilisé la mesure de qualité SSIM (Structural SIMilarity) [WBSS04] qui permet de déterminer la similarité structurelle entre les images de la vidéo tatouée et les images de la vidéo tatouée après l'application des différentes combinaisons d'attaques par un utilisateur. Cette mesure a été choisie puisqu'elle calcule la similarité de structure entre ces images et pas la différence pixel à pixel comme c'est le cas pour les autres critères tels que le PSNR partant de l'hypothèse que l'œil humain est plus sensible aux changements dans la structure de l'image. Cette étude de similarité sera appliquée après chaque validation de choix d'un utilisateur. En effet, si la valeur du SSIM est inférieure à 0.4 l'essai de l'utilisateur ne sera pas validé et les combinaisons qu'il a appliquées ne seront pas considérées.

5.1. Interaction des utilisateurs

Nous avons enregistré les différents choix d'attaques de chaque utilisateur afin de dégager les attaques les plus utilisées et plus importantes. Nous avons alors remarqué que les utilisateurs ont essayé de tester toutes les attaques afin de voir leurs impacts sur la vidéo. En fait, ils ont tous choisi les vidéos camcordées : 3 utilisateurs ont sélectionné la vidéo camcordée avec une caméra en face de l'écran, 9 ont sélectionné la vidéo camcordée avec une caméra à gauche de l'écran, 6 ont sélectionné la vidéo camcordée avec une caméra à droite de l'écran et 2 ont sélectionné la vidéo camcorder avec une caméra en dessous de l'écran. Pour la compression la plupart des utilisateurs (13 utilisateurs) ont choisi la compression MPEG-4 avec un débit de compression de 500 kbit/s afin d'éviter la dégradation de la qualité visuelle de la vidéo. La figure 7 présente pour chaque débit, le nombre d'utilisateurs qui l'ont choisi. Les utilisateurs qui ont choisi la vidéo camcordée avec une caméra à gauche, à droite ou en dessous de l'écran ont tous choisi d'appliquer la déformation afin de recadrer la vidéo. Enfin, pour les deux dernières attaques, les utilisateurs qui les ont choisi, ont bien veillé à ne pas trop dégrader la visibilité de la vidéo et surtout la visibilité de l'objet en mouvement Stefan. La figure 8 montre pour chaque attaque le nombre d'utilisateurs qui l'a choisi. D'après cette courbe, nous pouvons remarquer que les attaques les plus utilisées sont le camcording, la compression, la déformation et le cropping.

5.2. Robustesse des l'algorithmes de tatouage

L'algorithme de tatouage vidéo basé sur l'insertion multi-fréquentielle dans les régions d'intérêts [KJZ12], a montré

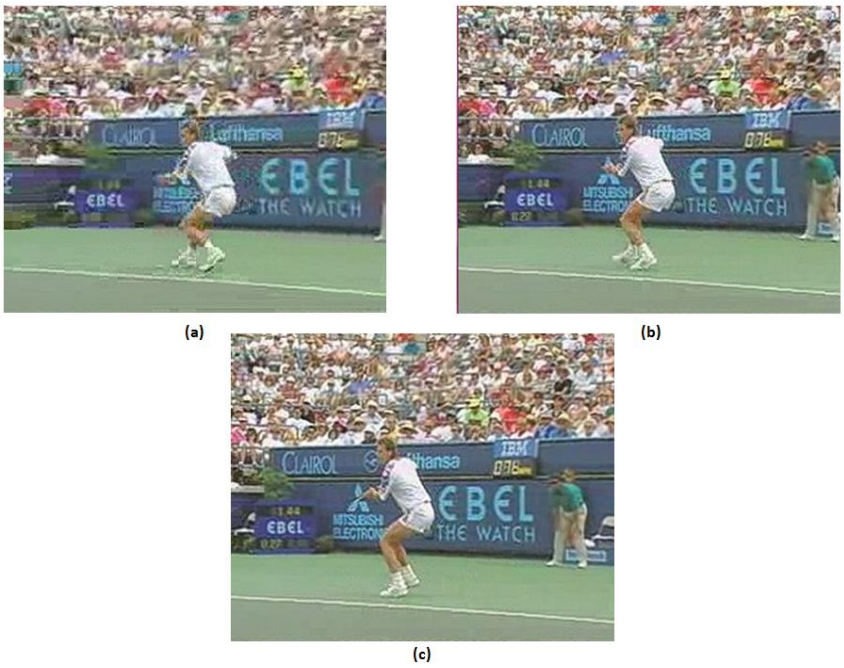


Figure 3: *Vidéo compressée : (a) Débit 200 kbit/s, (b) 500 kbit/s, (c) 1000 kbit/s.*

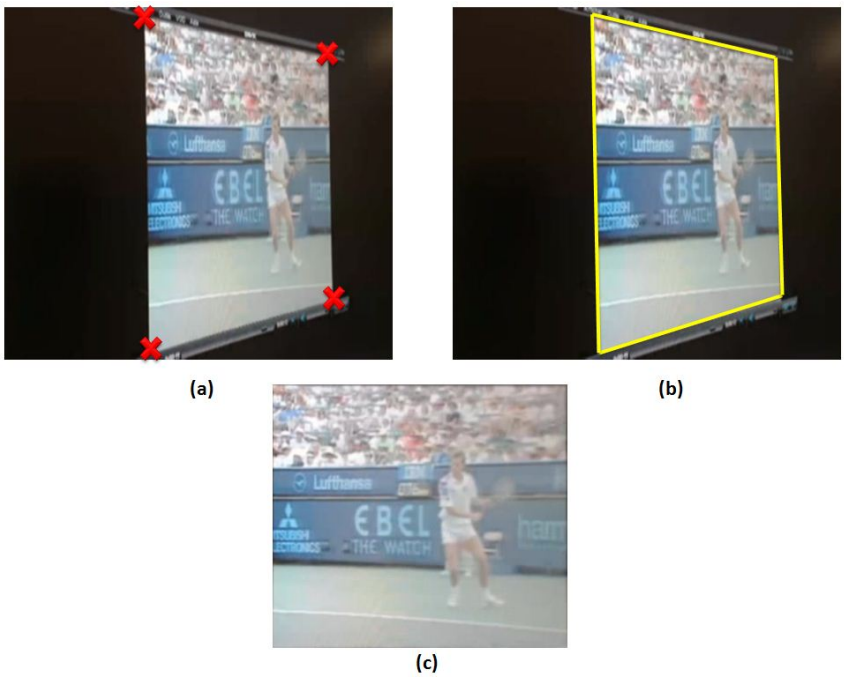


Figure 4: *Déformation des vidéos camcordées : (a) Sélection des coordonnées de la region à rectifier, (b) La région sélectionnée, (c) La région rectifiée.*

une haute performance face à toutes les combinaisons d'attaques possibles. En effet, on arrivait toujours à détecter la présence de la marque dans la vidéo après chaque test. La figure 9 présente le résultat de certaines combinaisons d'attaques appliquées sur la vidéo tatoué.

Pour l'algorithme de tatouage vidéo basé sur la transformée en ondelette [CL03], malgré sa robustesse face aux attaques usuelles, 6 utilisateurs ont réussi à détruire la marque insérée dans la vidéo tout en conservant une bonne visibilité, le tableau 1 présente les combinaisons appliquées par ces utilisateurs qui ont réussi à détruire la marque et les valeurs SSIM pour chaque combinaison et la figure 10 présente

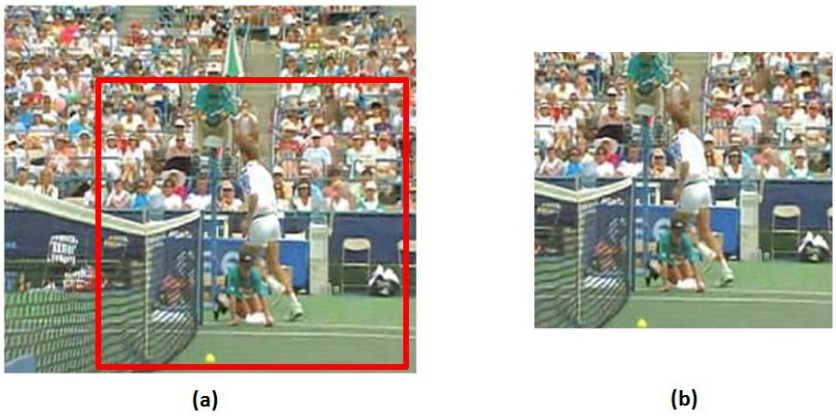


Figure 5: *Cropping de la vidéo : (a) sélection de la région par l'utilisateur, (b) vidéo croppée.*

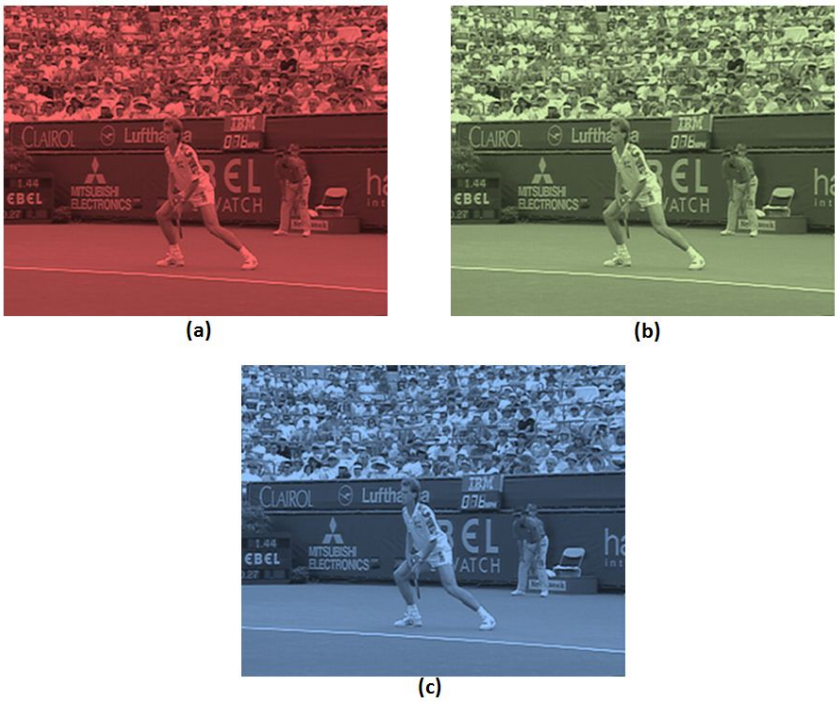


Figure 6: *Modification des couleurs de la vidéo : (a) Ajout de rouge, (b) Ajout de vert, (c) Ajout de bleu.*

le résultat de deux combinaisons d'attaques appliquées sur la vidéo tatoué.

6. Conclusion

Dans ce papier nous avons proposé un nouveau jeu d'attaques basé sur le crowdsourcing dans le but d'améliorer l'évaluation des algorithmes de tatouage vidéo. Ce jeu permet aux différents utilisateurs d'appliquer sur une vidéo tatouée des combinaisons d'attaques importantes et réelles telle que le camcording, la déformation, l'ajout de couleur et la compression dans le but de détruire la marque qui a été insérée. Ce jeu a permis non seulement d'évaluer les algorithmes d'insertion utilisés pour tatouer la vidéo test, mais aussi de dégager l'ensemble d'attaque les plus importantes en se basant sur les choix des différents utilisateurs. En effet, les tests réalisés ont montré que notre méthode d'évaluation est bien meilleure que les autres méthodes usuelles vue que ça nous a permis de comparer deux algorithmes de tatouage vidéo qui ont déjà fait leurs preuves face aux attaques usuelles et ont été classés parmis les méthodes robuste de tataouge vidéo alors que un seul a résisté aux combinaisons d'attaques appliquées dans notre jeu. En plus, ça a montré l'importance de l'attaque du camcording qui présente un grand risque pour les algorithmes de tatouage et qui est souvent négligée dans l'évaluation de robustesse de la plupart des algorithmes de tatouage vidéo. Comme perspective pour ce travail, nous allons le compléter afin d'élaborer un nouveau protocole d'évaluation des techniques de tatouage vi-

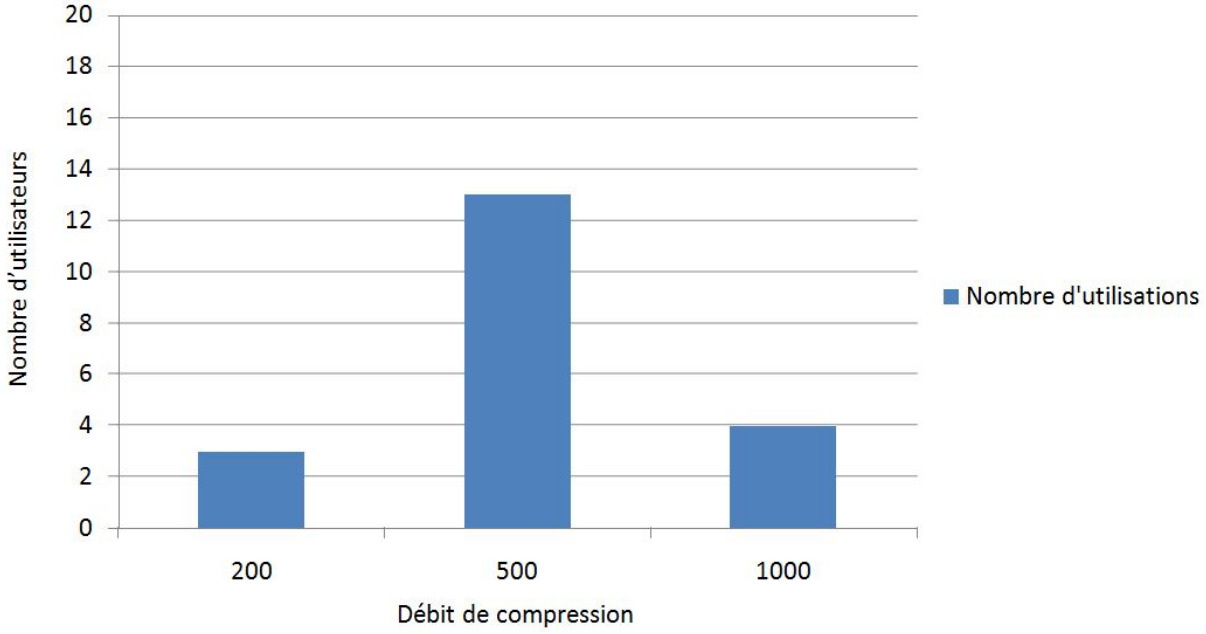


Figure 7: *Nombre d'utilisation pour chaque débit de compression.*

	(1)	(2)	(3)	(4)	(5)	(6)
SSIM	0.57	0.51	0.53	0.437	0.431	0.47

Table 1: *Valeurs de SSIM pour les vidéos tatouées dont la marque à été détruite : (1) Vidéo camcordée en face + compression 500 kbit/s + Cropping,(2) Vidéo camcordée à droite + déformation + compression 500 kbit/s + Cropping,(3) Vidéo camcordée à gauche + déformation + compression 500 kbit/s,(4) Vidéo camcordée à gauche + déformation + compression 500 kbit/s + Cropping + modification de couleurs,(5) Vidéo camcordée à gauche + déformation + compression 500 kbit/s + Cropping,(6) Vidéo camcordée en bas + déformation + compression 500 kbit/s + Cropping*

déo et ceci en ajoutant d'autres attaques qui peuvent présenter un danger sur la marque en la combinant avec d'autres attaques et aussi en intégrant l'évaluation de la visibilité de la vidéo après l'application des attaques. Ce jeu sera bénéfique sur deux côtés. En effet, il permettra d'évaluer l'algorithme de tatouage à tester en se basant sur un ensemble d'attaques les plus importantes. En plus, l'étude des choix des différents utilisateurs ainsi que la réaction de l'algorithme de tatouage face aux tentatives de sa destruction pourra servir pour l'amélioration du principe de l'algorithme afin de résister à ces attaques.

Références

[A.P99] A.P.PETITCOLAS F. : Attaques et évaluation des filigranes numériques. *CORSEA*, Num. 14-15 (1999).

[BF] BAS P., FURON T. : Project bows2. *www.http ://bows2.ec-lille.fr/.*

[C.08] C. B. D. : Crowdsourcing as a model for problem solving : an introduction and cases. *The International Journal of Research into New Media Technologies. Vol. 14*, Num. 1 (2008).

[C.10] C. B. D. : Moving the crowd at threadless : motivations for participation in a crowdsourcing application. *Information, Communication & Society. Vol. 13*, Num. 8 (2010).

[CCOM10] CARLIER A., CHARVILLATA V., OOI W., MORIN. R. G. G. : Crowdsourced automatic zoom and scroll for video retargeting. *ACM Multimedia* (2010).

[CL03] CHAN P. P.-W., LYU M. R. : A dwt-based digital video watermarking scheme with error correcting code. *ICICS'03* (2003).

[CMR01] CHANDRAMOULI R., MEMON N., RABBANI M. : Invisible watermarking for digital cinema. *DIGITAL WATERMARKING* (2001).

[FJB99] F.HARTUNG, J.K.SU, B.GIROD : Spread spectrum watermarking : Malcious attacks and counter-attacks. *SPIE : Security and watermarking of Multimedia Contents. Vol. 3657*, Num. 147-158 (janvier 1999).

[GF11] GREENGARD, FOLLOWING S. : the crowd. *Communications of the ACM. Vol. 54*, Num. 2 (2011).

[J.06] J. H. : The rise of crowdsourcing. *Wired Magazine. Vol. 14*, Num. 6 (2006).

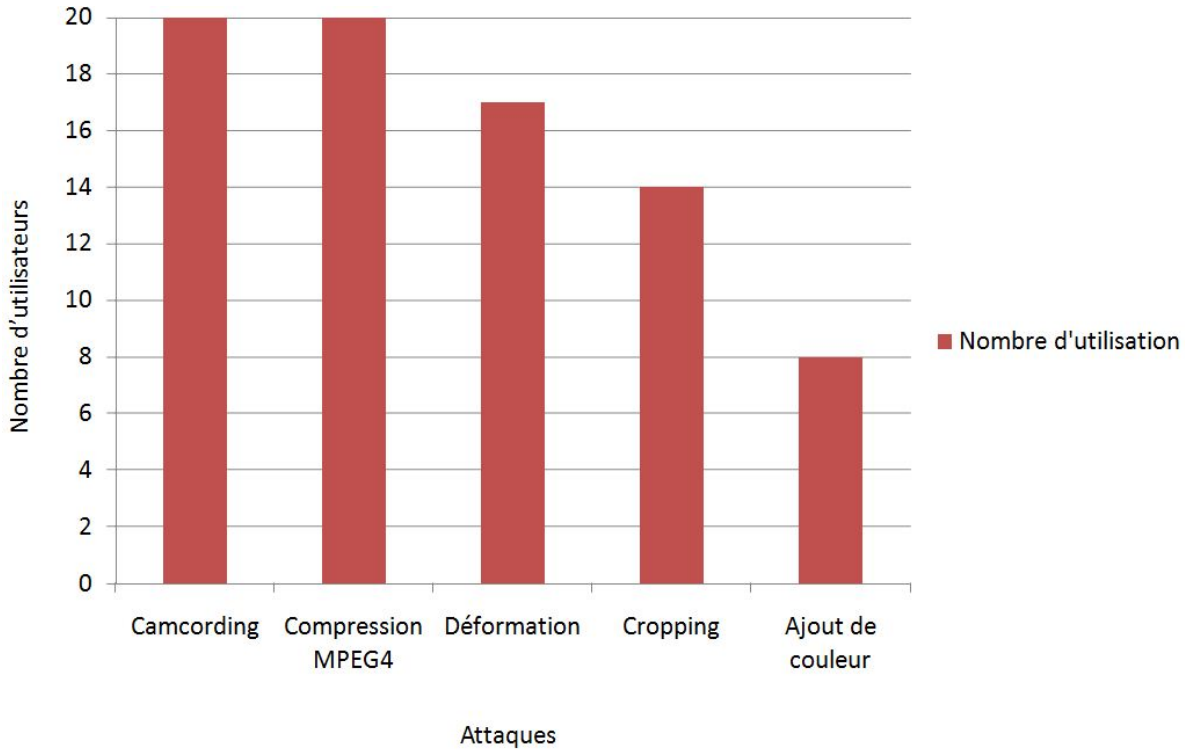


Figure 8: *Nombre d'utilisation pour chaque attaque*

Figure 9: *Vidéo attaquée, tatoué par l'algorithme [KJZ12] :(b) vidéo camcorder en face + compression MPEG-4 200kbit/s + cropping, (c) vidéo camcorder à gauche + compression 500 kbit/s + cropping + ajout de couleur bleu +10.*

[J.08] J. B. : Hack, mash and peer : Crowdsourcing government transparency. *The Columbia Science and Technology Law Review. Vol. 9* (2008).

[KJZ12] KERBICHE A., JABRA S. B., ZAGROUBA E. : A robust video watermarking based on image mosacing and multi-frequential embedding. *IEEE International Conference on Intelligent Computer Communication and Processing* (2012).

[Pet00] PETITCOLAS. F. A. P. : Watermarking schemes evaluation. *I.E.E.E. Signal Processing. Vol. 17*, Num. 5 (2000).

[Rol] ROLLIN C. : Certimark. *www.certimark.org/*.

[SKWE14] SCHABER P., KOPF S., WESCH C., EFFELSBERG W. : A camcorder copy simulation as watermarking benchmark for digital video. *ACM Multimedia Systems Conference* (2014).

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SI-

Figure 10: *Vidéo attaquée, tatoué par l'algorithme [CL03] : (a) Vidéo camcordée à gauche + déformation + compression 500 kbit/s(b) Vidéo camcordée à gauche + déformation + compression 500 kbit/s + Cropping*

MONCELLI E. P. : Image quality assessment : From error measurement to structural similarity. *IEEE Trans. Image Processing* (2004).

[XLGyM05] XIE X., LIU H., GOUMAZ S., YING MA. W. : Learning user interest for image browsing on small-formfactor devices. *SIGCHI Conference on Human Factors in Computing Systems* (2005).

Schéma conjoint de tatouage et compression des LDI(s) pour des flux auto-stéréoscopiques

N. Khelfi[1], Z. Ahmed Foitih[2] et L. Lucas[3]

[1]LRIIR, Faculté des Sciences Exactes et Appliquées Université d'Oran - Algérie
[2]Département d'Electronique - Faculté du Génie Electrique USTO Mohamed Boudiaf - Oran - Algérie
[3]CReSTIC SIC EA3804 - Université de Reims Champagne-Ardenne

Résumé

Dans cet article est présenté un schéma conjoint de tatouage et compression des LDI(s) (Layer Depht Image) générées à partir d'un jeu d'images multi-vues. Chaque calque ou "layer" correspond à une vue et contient les informations qui ne sont pas contenues dans les autres calques. Les calques de profondeur sur lesquels nos travaux se basent contiennent uniquement des informations non redondantes (ou résiduelles). Elles sont extraites à partir de n vues d'entrée et de leur carte de disparité et correspondent à des zones occultées. Les pixels occlus, qui sont susceptibles d'être visualisés à partir d'autres angles de vue, se voient conférer une plus forte protection contre les erreurs de transmission ou de compression. Nous cherchons donc à effectuer l'insertion d'une marque dans un ensemble de LDI(s) pendant la phase de quantification et de compression à l'aide d'une méthode qui doit tenir compte de ces pixels occlus (ou résiduels) localisés dans des zones dispersées de chaque layer. Notre approche de tatouage joint à la compression est basée sur la quantification vectorielle algébrique à zone morte (QVAZM). Cette technique permet d'effectuer conjointement compression et tatouage à l'aide de la QVAZM en réalisant simultanément la quantification et l'insertion de la marque durant la compression.

Mots clé : Images auto-stéréoscopiques, Cartes de profondeur, Layer Depth Image, Sécurité, Tatouage numérique, Quantification vectorielle à zone morte, Compression

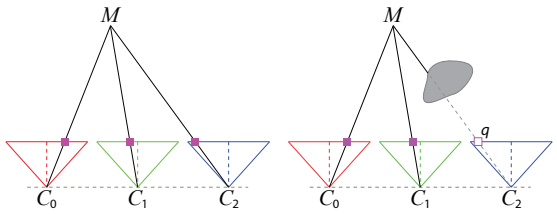
1. Introduction

Le nombre croissant d'applications utilisant les technologies multimédias numériques a fortement accentué la nécessité d'assurer la protection de ces médias notamment en termes de droit d'auteur. Nous nous intéressons ici à l'une de ces techniques, le tatouage numérique d'images ou de vidéos, qui permet d'insérer de manière imperceptible par l'utilisateur final un message (copyright, message de vérification, ...) à un fichier hôte. Ce message, généralement appelé marque, se présente comme un ensemble de bits dont le contenu dépend de l'application. De nombreuses méthodes d'insertion existent dans la littérature [CMB*08a]. De manière générale, la détection de la marque doit être fiable, sans fausse détection et si possible pas de faux rejet. En outre, les marques doivent être :

- sécurisées - une modification non autorisée comme le retrait de la marque doit être impossible ;
- insensibles - il doit être possible de détecter et de décoder la marque sans l'utilisation de l'image ou de la vidéo d'origine ;
- perceptuellement invisibles et ne doivent pas dégrader la qualité d'image ou de vidéo ;
- résistantes aux manipulations telles que les transformations photométriques (*e.g.* correction de filtrage ou luminance), transformations géométriques (*e.g.* translation, rotation, redimensionnement, recadrage ou changement aspect ratio), conversions analogiques-numériques (resp. numériques-analogiques), la numérisation, la compression avec perte et autres attaques cryptographiques.

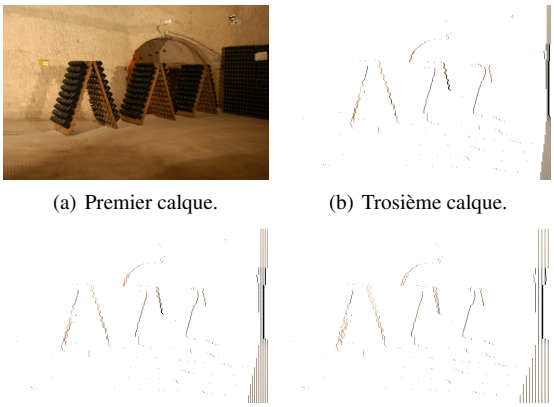
Dans le cadre de la vidéo et de surcroit de la vidéo 3D [KCA10, MCP13, BBD14], la marque doit résister à différents schémas de compression vidéo, ne pas altérer l'information de profondeur et être détectable n'importe où dans le film y compris sur un court laps de temps.

Ce article présente une idée de schéma conjoint de tatouage et de compression sur des images type LDIs, générées à partir d'un jeu d'images auto stéréoscopiques, prises à partir de n points de vue ($n = 8$). Notre travail s'intègre dans un processus élaboré par [Bat12, Niq11]. Aussi, dans la première partie de cet article, seront définies les LDI(s) et les différentes étapes de traitement qui leur ont été appliqué et qui rentrent dans le processus de compression. C'est sur ces LDIs, particulièrement les calques résiduels que va être appliqué un schéma d'insertion d'une marque. Cette étape d'insertion s'applique au moment de la compression des calques. Après avoir subi une transformation en ondelettes, ces images seront quantifiées par la méthode de Quan-



(a) Sans occultation. (b) Avec occultation.

Figure 1: *Soit M un point 3D et son match correspondant m (■ pixel ∈ m, □ pixel ∉ m). Sur la figure (b), le point M est caché par l'objet gris et n'est pas visible depuis la vue 2. Par conséquent, le pixel $q = (2, m_x - 2m_d, m_y)$ dans la vue 2 ne correspond pas à la projection de M et ne fait donc pas partie de m même s'il satisfait à l'équation 2.1. Ainsi, nous dirons que M est visible dans les vues 0 et 1 et non visible dans la vue 2.*



(a) Premier calque. (b) Trosième calque.

(c) Cinquième calque. (d) Septième calque.

Figure 2: *Calques générés à partir de la séquence "Ruinart".*

tification Vectorielle Algébrique à Zone Morte qui permet de localiser les zones de forte activité représentées par les vecteurs d'ondelettes dits significatifs et qui sont de bons candidats au tatouage. Cela permet d'améliorer la robustesse du tatouage. L'insertion de la marque se fait donc sur ces vecteurs significatifs, les vecteurs seuils (ou nuls) sont exclus du processus de tatouage.

2. Génération de LDI

2.1. Estimation des disparités

La plupart des problèmes rencontrés par les méthodes de compression de vidéo 3D basées LDV (Layer-Depth-Video) sont dus à des incohérences entre les cartes de disparité le plus souvent présentes dans les zones d'occlusion. Pour remédier à cela, nous nous basons dans cet article, sur des algorithmes d'estimation de disparité fonctionnant en arithmétique entière de sorte à ce que tous pixels correspondants aient la même ordonnée dans leur base d'image respective. L'approche retenue est celle de [NPR10a] qui permet à la fois de produire des disparités cohérentes et de lever toutes ambiguïtés.

Soit P l'ensemble des pixels contenus dans les n vues. On définit $p = (i, p_x, p_y)$ comme le pixel de coordonnées (p_x, p_y) de la vue i et m comme un match (un ensemble de pixels dont les projections ont pour origine le même point 3D M - voir Figure 1(a)). Soit (x_0, y_0), les coordonnées du pixel $\in m$ dans la première vue (la vue 0). Dans le cadre d'une géométrie de capture simplifiée (distribution linéaire équidistantes des centres optiques), les coordonnées des autres pixels de m dans la base induite par les autres vues sont alors faciles à trouver. Par exemple, la position du pixel dans la vue k est donnée par

$$\begin{pmatrix} x_k \\ y_k \end{pmatrix} = \begin{pmatrix} x_0 - k \times m_d \\ y_0 \end{pmatrix},$$

où m_d est une valeur de disparité entière associée à m.

L'utilisation de disparité entière assure que l'expression $x_0 - k \times m_d$ correspond à une coordonnée entière, ce qui évite tous problèmes d'ambiguïté dans le processus d'extraction des différents calques de disparité. Par convention, nous posons $m_x = x_0$ et $m_y = y_0$, définissant la position de la projection de M dans la première vue (la vue 0). Ainsi, un pixel p de la vue k peut appartenir à un match m si et seulement si la condition suivante est vraie :

$$\begin{cases} p_x &= m_x - k \times m_d \\ p_y &= m_y \end{cases}$$

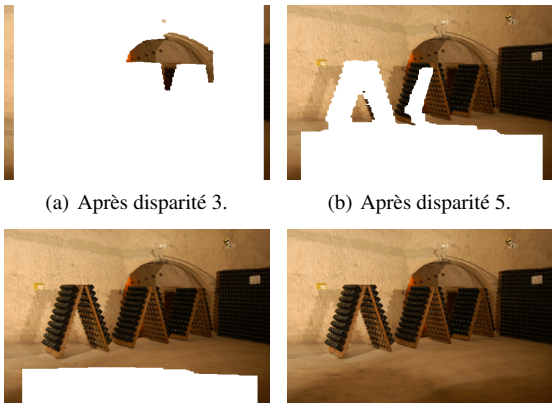
L'algorithme d'estimation de disparité vise à déterminer une partition de tous les pixels dans les matchs. Cette partition est un ensemble de correspondances non vides (contenant chacun au moins un pixel) qui n'ont aucun pixel en commun. De plus, l'union de tous les matchs de la partition est égale à l'ensemble P de tous les pixels. Une partition g vérifie donc les conditions suivantes :

$$\begin{aligned} &\forall m \in g, m \neq \emptyset \\ &\forall m, n \in g, m \cap n = \emptyset \\ &\textstyle\bigcup_{m \in g} m = P \end{aligned}$$

Dans la pratique, cet algorithme commence par initialiser la partition avec les matchs correspondant aux plus grandes valeurs de disparité. Toutes ces valeurs de disparité sont ensuite analysées des plus élevées aux plus petites. Pour chaque valeur de disparité α, tous les matches déjà présents dans la partition sont pris en compte et nous cherchons à évaluer s'ils existent vraiment ou pas. Si un match n'existe pas, il est alors retiré de la partition et remplacé par de nouveaux matches de disparité α' plus faibles (a priori donc visibles). L'ensemble de ces contraintes sont globalement maintenues cohérentes au moyen d'un algorithme de graph-cuts [BVZ99, KZ04]. Pour plus de détails sur cette approche, il est possible de se référer à [NPR10b].

2.2. Extraction des calques de disparités

Le principe de mise en correspondance présenté précédemment permet de détecter l'ensemble des redondances. Tous les pixels appartenant à un même match sont des projections d'un point 3D unique de la scène et sont considérés comme redondants. Supprimer ces redondances revient donc à ne maintenir qu'un seul pixel de chaque match de la partition (celui ayant le plus petit ordre dans l'image multi-vues).



(a) Après disparité 3. (b) Après disparité 5.

(c) Après disparité 7. (d) Après disparité 10.

Figure 3: *Reconstruction de l'image à partir de calques.*

Ainsi, le premier calque conservera tous les pixels, le second ne conservera que les pixels ne pouvant être déduits du premier calque, le troisième conservera uniquement les pixels ne pouvant être déduits ni du premier ni du second calque et ainsi de suite. Notre définition du LDI diffère de celle habituellement décrite dans la littérature [YLKYS07, JMG09] : dans notre cas, le nombre de calques est toujours égal à n et chacun est attaché à sa vue correspondante (et pas à une vue de référence unique). En outre, le rapport de pixels est approximativement constant d'un calque à l'autre, à l'exception du premier, bien entendu. Afin de réduire les artefacts dus aux variations de lumière d'une image à l'autre, seule la couleur moyenne de tous les pixels d'un match est conservée. La figure 2 illustre ce procédé d'extraction des différents calques sur la séquence "Ruinart".

2.3. Reconstruction de l'image

Le processus de reconstruction des différentes vues consiste s'appuie pour partie sur l'algorithme du peintre [NNS72]. Il exploite également la contrainte exposée précédemment, contrainte de cohérence qui assure qu'un match ne peut être visible dans une autre image que si celui-ci est obstrué par un autre match avec un écart supérieur. L'idée consiste donc à "étaler" les pixels par ordre de disparité croissante dans les n. Certains pixels pourront être dessinés sur d'autres déjà existants de sorte à reproduire les zones occlusions présentes dans la scène. La figure 3 illustre plusieurs étapes de cette phase de reconstruction d'une image à partir des calques de la Figure 2.

3. Compression des LDI(s)

Nous avons présenté dans la précédente section une structure de représentation d'images multi-vues appelée LDI. Cette structure permet au moyen de calques de stocker pour chaque pixel non seulement sa couleur mais également sa profondeur relativement au calque qui le porte. Elle est générée à partir d'un jeu d'images multi-vues. Les n calques (ou "layers") sont associées à chacune des n vues. La première couche, appelée couche de référence, contient tous les pixels de la première vue, alors que les $n-1$ couches suivantes (désignées par le terme "couches résiduelles") contiennent uniquement l'information n'appartenant pas à la première couche (à cause des zones d'occultations).
La compression d'images multi-vues [Bat12] basée sur ce principe repose sur les étapes suivantes :

- Etape 1 - Conversion colorimétrique de l'espace RGB vers l'espace YCbCr.
- Etape 2 - Application de la DWT sur chacune des $n-1$ calques : ces calques ne contenant que des pixels résiduels localisés dans des zones dispersées, la SA-DWT (Shape-Adaptative Discrete Wavelet Transform) [LL00] a été utilisée pour ne compresser que certaines zones d'une image tout en omettant le reste de l'information. La sélection de ces zones a été faite par masque binaire lui indiquant si le pixel est à prendre en considération pour la compression ou pas.
- Etape 3 - Codage des différents coefficients d'ondelettes par application de l'algorithme SPIHT associé à un codeur arithmétique binaire.

Notre objectif est de présenter un schéma de tatouage conjoint à la compression pour insérer une marque dans ces calques résiduelles en choisissant une méthode qui tienne compte des particularités de ces LDIs. La méthode qui nous a semblé la plus appropriée à ce type d'images est celle qui utilise la quantification vectorielle algébrique à zone morte (QVAZM) et d'insérer la signature sur des coefficients quantifiés de façon à répartir les bits de la signature sur les différents calques.

4. Le tatouage joint à la compression

La compression jointe au tatouage a suscité un intérêt récent. Associer tatouage et compression permet d'améliorer la capacité d'insertion du tatouage et d'assurer une meilleure détection tout en maintenant une bonne qualité d'image. La compression représente, non seulement un passage obligé du stockage ou du transfert d'images, mais aussi l'une des attaques les plus destructrices vis-à-vis du tatouage.

Le domaine multi-résolution est la base des normes de compression avec perte. Son utilisation comme support d'insertion de la signature assure une meilleure robustesse vis à vis de cette norme de compression. L'ondelette 9/7 "Daubechies" utilisée, permet le passage du domaine spatial vers le domaine multi-résolution.

Lorsque la marque est insérée directement durant le processus de codage, celle-ci devient partie intégrante de l'image. Pour ce faire, nous intégrons l'étape d'insertion dans le schéma de codage basé sur la quantification vectorielle algébrique avec zone morte afin d'en exploiter les principaux avantages, à commencer par son aptitude à localiser les zones de forte activité. Ces zones sont représentées par les vecteurs d'ondelettes dits "significatifs" qui sont donc de bons candidats au tatouage. Aussi, la phase d'insertion du message se situe pendant l'étape 2 de l'algorithme de compression sus cité (cf. section 3) à partir de l'application de la DWT sur les $n-1$ calques résiduels. L'application du masque binaire et de la SA-DWT sur les calques laisse place à la quantification algébrique à zone morte. Les quelques vecteurs quantifiés sélectionnés seront alors tatoués.

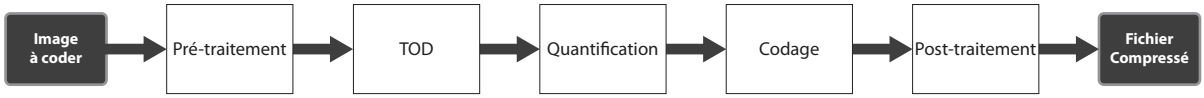


Figure 4: *Schéma général de la chaine de compression d'une image.*

Le tatouage d'image est divisé en deux parties : la première est la phase d'insertion de la marque, la deuxième est la phase de détection. Nous ne considèrerons, dans cet article, que la phase d'insertion.

4.1. Méthode de tatouage proposée

Le schéma de tatouage que nous proposons se base sur un schéma substitutif dans le domaine multi-résolution où les coefficients d'ondelettes ont subi une Quantification Vectorielle Algébrique à Zone Morte (QVAZM) [Voi03]. L'image obtenue est donc une image comprimée et tatouée. En d'autres termes, nous appliquons notre technique de tatouage directement sur les données quantifiées. Différents travaux ont montré l'intérêt d'utiliser ce type de schéma [KH98, CMB*08b]. L'objectif commun de la compression et du tatouage d'images est de minimiser l'impact visuel des modifications apportées à l'image originale. Cependant, les modifications apportées ont des objectifs totalement opposés : la compression d'images vise à supprimer les données imperceptibles afin de réduire la quantité d'information, alors que le tatouage consiste à ajouter une information imperceptible (intégration d'une signature).

Ce schéma de tatouage (cf. Figure 5) sera appliqué sur l'ensemble des calques résiduels contenant des pixels correspondants aux zones d'occlusions. La particularité de ce schéma est que la marque (ou le message) sera étendu sur les différentes couches résiduelles. De ce fait, la première étape consiste à sélectionner les vecteurs de coefficients d'une seule sous bande par couche qui vont porter une partie de la signature et la deuxième permet l'insertion de cette dernière sur les coefficients de ces sous bandes de marquage suivant une fonction d'insertion donnée.

La justification de ce schéma de tatouage se base sur aspects suivants :

– Le domaine multi-résolution : c'est le domaine le mieux adapté au système visuel humain, dans le sens où celui-ci est plus ou moins sensible à certaines gammes fréquentielles sélectionnées. La décomposition en sous-bandes de la transformée en ondelettes permet d'isoler les fréquences adaptées au tatouage. De plus, c'est le domaine utilisé par le standard de compression SPIHT déjà utilisé dans notre cas. Le niveau de décomposition représentant le niveau d'insertion est choisi de telle sorte que les coefficients de ce niveau soient généralement bien préservés lors de la phase de quantification. Généralement, c'est ceux qui contiennent les informations les plus significatives de l'image, cela permet de mieux préserver la signature face à différentes attaques. De plus la modification de certains coefficients de ce niveau n'engendre pas de dégradations trop importantes lors de la phase de reconstruction de l'image. Ce niveau se situe entre les basses et moyennes fréquences.
– Un schéma substitutif : ce type de schéma, contrairement au schéma additif, permet d'utiliser au maximum la capacité du support d'insertion. Par exemple, il est possible d'insérer 1 bit d'information issu du message par vecteur (non nul) d'une sous-bande d'un calque.
– La QVAZM : la Quantification Vectorielle Algébrique avec Zone Morte d'une sous-bande présente l'avantage de ne conserver que les coefficients d'ondelettes significatifs, les autres étant mis à zéro. Elle permet ainsi d'améliorer la robustesse de la signature insérée face à différents types d'attaques [LL00]. Le choix de la QVAZM, dans ce cas, est justifié d'une part par le fait que les couches de la LDI contiennent peu d'informations correspondants aux pixels occlus et que ces informations sont dispersées et d'autre part parce que les différentes sous bandes sont considérées indépendantes les unes des autres et de ce fait, la modification de certaines d'entre elles, par insertion d'une marque, ne perturbe pas les autres (approche intra bande). En plus, le choix des sous bandes peut être différent d'un calque à l'autre.

4.2. L'opération d'insertion

L'opération d'insertion se fait de la même manière sur les différents calques d'une LDI à l'exception d'une sélection différente des vecteurs à tatouer. Elle consiste en fait à sélectionner des vecteurs des sous bandes puis à insérer les bits de la signature sur les différents vecteurs des différentes couches de la LDI.
Les étapes d'insertion sont présentées pour un calque de la LDI et doivent être répétées sur les autres calques :

– La phase de Quantification. Le schéma de quantification est réalisé comme suit : une sous bande SB_k est sélectionnée. De cette sous bande sont sélectionnés les vecteurs susceptibles d'être tatoués et vont donc être quantifiés par la QVAZM : (vecteurs quantifiés Y_i).
– La phase de tatouage : A l'aide de la clé K, vont être sélectionnés les vecteurs quantifiés qui vont porter la signature S. Le tatouage est réalisé par un schéma de substitution. Il permet d'insérer 1 ou quelques bits de la signature $w = (b_0, \cdots, b_{M_w})$ sur des vecteurs quantifiés Y_i de la sous-bande SB_k. A l'aide de la clé K, vont être sélectionnés les vecteurs $S(Y_i, K)$ qui vont porter la signature. L'ensemble de ces vecteurs est noté : $S(Y_i, K) = Y_s$.

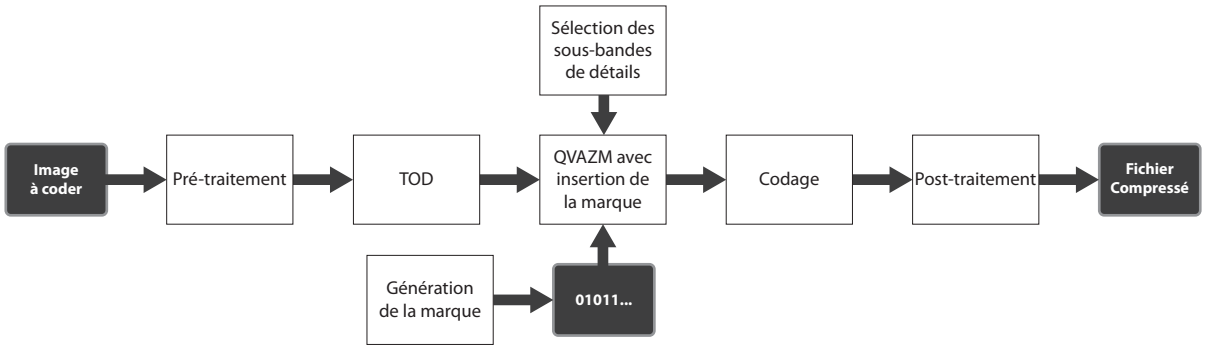


Figure 5: *Schéma général du tatouage d'une image joint à la compression.*

Les vecteurs qui n'ont pas été sélectionnés sont notés $Y_{\widetilde{S}}$. Nous avons donc les propriétés suivantes.

$$Y_i = Y_s \cup Y_{\widetilde{S}}$$

$$Y_s \cap Y_{\widetilde{S}} = 0$$

Si l'on considère qu'un bit b_j de la signature est inséré par vecteur sélectionné Y_s alors ce vecteur tatoué s'écrira : $W_b(Y_{s_j}, b_j) = Y_{s_j}^w$ (vecteur sélectionné tatoué avec le bit b_j de la marque).

L'étape d'insertion modifie les composantes du vecteur Y_{S_j} en fonction du bit b_j à insérer. Appliquée à tous les vecteurs sélectionnés, l'étape d'insertion conduit à l'ensemble des vecteurs tatoués Y_s^w.

$$W_s(Y_s, b) = Y_s^w \text{ avec } Y_s^w = \{Y_{s_j}^w\}$$

Les vecteurs sélectionnés et tatoués Y_s^w vont être ensuite intégrés aux vecteurs non sélectionnés $Y_{\widetilde{S}}$ pour former l'ensemble des vecteurs Y_w de la sous-bande tatouée SB_k^w. Ces différentes étapes réalisent la substitution $Sub(Y_s, Y_s^w, K)$ des vecteurs avec les vecteurs tatoués Y_s^w afin d'intégrer la signature. Il suffit ensuite de renouveler ces différentes étapes sur d'autres sous-bandes et sur les autres calques de la LDI.

5. Conclusion

Nous avons présenté dans cet article le principe d'un schéma de tatouage d'images multi-vues représentées par une structure LDI. Les calques de cette structure appelés calques résiduels ne contiennent que peu d'informations reparties d'une façon très disperses. L'étape de tatouage à été jointe à la quantification vectorielle algébrique à zone morte qui permet de ne conserver que les vecteurs des coefficients qui correspondent à la description des contours (détails) et d'éliminer les vecteurs des coefficients de valeurs nulles qui correspondent à des zones homogènes. Le schéma de tatouage doit être appliqué à l'ensemble des calques de la LDI avec la répartition des bits du message sur l'ensemble des calques avec un choix adéquat des sous bandes pour chaque calque. Ce travail est encore au stade expérimental. Le développement est en cours de réaliser et les premiers tests devraient être réalisés prochainement.

Références

[Bat12] BATTIN B. : *Compression multi-vues de flux autostéréoscopiques*. PhD thesis, 2012. Thèse de doctorat dirigée par Lucas, Laurent Informatique Reims 2012.

[BBD14] BURINI C., BAUDRY S., DOËRR G. : Blind detection for disparity-coherent stereo video watermarking, 2014.

[BVZ99] BOYKOV Y., VEKSLER O., ZABIH R. : Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 23* (1999), 1222–1239.

[CMB*08a] COX I., MILLER M., BLOOM J., FRIDRICH J., KALKER T. : *Digital Watermarking and Steganography*, 2 ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[CMB*08b] COX I., MILLER M., BLOOM J., FRIDRICH J., KALKER T. : *Digital Watermarking and Steganography*, 2 ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[JMG09] JANTET V., MORIN L., GUILLEMOT C. : Incremental-ldi from multi-view coding. In *3DTV-Conference 2009 : The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON 2009)* (2009).

[KCA10] KOZ A., CIGLA C., ALATAN A. : Watermarking of free-view video. *Image Processing, IEEE Transactions on. Vol. 19*, Num. 7 (July 2010), 1785–1797.

[KH98] KUNDUR D., HATZINAKOS D. : Digital watermarking using multiresolution wavelet decomposition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Seattle, Washington, USA, May 12-15, 1998* (1998), pp. 2969–2972.

[KZ04] KOLMOGOROV V., ZABIH R. : What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 26* (2004), 147–159.

[LL00] LI S., LI W. : Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding.

IEEE Transactions on Circuits and Systems for Video Technology. Vol. 10, Num. 5 (Aug 2000), 725–743.

[MCP13] MITREA M., CHAMMEM A., PRÊTEUX F. : Tatouage stéréoscopique. In *Vidéo 3D - Capture, traitement et diffusion*, Lucas L., Loscos C., Remion Y., (Eds.). Hermès - Lavoisier, septembre 2013, pp. 251–272.

[Niq11] NIQUIN C. : *Reconstruction du relief et mixage réel/virtuel par caméras relief multi-points de vues*. PhD thesis, 2011. Thèse de doctorat dirigée par Remion, Yannick Informatique Reims 2011.

[NNS72] NEWELL M. E., NEWELL R. G., SANCHA T. L. : A new approach on the shaded picture problem. *Proceedings of the ACM National Conference* (1972), 443 – 450.

[NPR10a] NIQUIN C., PRÉVOST S., REMION Y. : An occlusion approach with consistency constraint for multiscopic depth extraction. *International Journal of Digital Multimedia Broadcasting* (2010).

[NPR10b] NIQUIN C., PRÉVOST S., REMION Y. : A point cloud based pipeline for depth reconstruction from autostereoscopic sets, 2010.

[Voi03] VOINSON T. : *Quantification vectorielle algébrique avec zone morte : application à la compression d'images à bas débit et au tatouage d'images*. 2003.

[YLKYS07] YOON S.-U., LEE E.-K., KIM S.-Y., YO-SUNG H. : A framework for representation and processing of multi-view video using the concept of layered depth image. *Journal of VLSI Signal Processing. Vol. 46* (2007), 87–102.

Frédéric Dufaux

Frédéric Dufaux est Directeur de Recherche CNRS à Télécom ParisTech depuis 2010. Il est également rédacteur en chef de la revue scientifique *Signal Processing: Image Communication* publiée par Elsevier. Frédéric a obtenu son master en physique et son doctorat en sciences techniques de l'EPFL en 1990 et 1994 respectivement. Frédéric a plus de 20 ans d'expérience dans le domaine de la recherche, occupant des postes à l'EPFL, MIT, Digital Equipment Corp. et Compaq Computer Corp. Il est impliqué dans la normalisation des technologies de vidéo et d'imagerie numérique. Il a été co-chair de JPEG 2000 Wireless (JPWL) et co-chair de JPSearch. Il est le récipiendaire de deux prix de l'ISO pour ces contributions. Frédéric était le *Vice General Chair* de la conférence IEEE *Int. Conf. on Image Processing* 2014. Il est un membre élu des Comités Techniques *Image, Video, and Multidimensional Signal Processing* (IVMSP) et *Multimedia Signal Processing* (MMSP) de l'IEEE.Ses travaux de recherche portent sur le codage vidéo, la vidéo 3D, le *High Dynamic Range*, l'évaluation de la qualité visuelle, la vidéo surveillance et l'analyse vidéo. Il est auteur ou co-auteur de plus de 100 publications. Il est coéditeur de l'ouvrage *Emerging Technologies for 3D Video – Creation, Coding, Transmission and Rendering*. Il détient 17 brevets délivrés ou en attente.

Vidéo 3D – technologies existantes et émergentes

La vidéo en relief ou 3D offre la perspective d'une amélioration notable de la qualité et de l'expérience perçue par l'utilisateur, et est largement considérée comme une innovation majeure dans le domaine de la vidéo numérique. Néanmoins, cette nouvelle technologie soulève de nombreux sujets de recherche, concernant l'acquisition du contenu, sa représentation, sa compression et son affichage. Dans cet exposé, je vais premièrement aborder les technologies existantes pour la vidéo 3D. Je vais également discuter leurs limitations. En effet, les représentations actuelles – stéréoscopique ou multi-vues – n'exploitent qu'un aspect de la perception de la profondeur. Dans un deuxième temps, je vais présenter de nouvelles directions de recherche émergeantes, incluant le Super Multi-Vues et l'holographie. En particulier, l'holographie offre tous les indices de profondeur utilisés par l'être humain, avec le potentiel d'être l'expérience 3D suprême.

Squelette Euclidien Discret Connecté (DECS) résistant au bruit pour l'appariement de formes basé graphes

A. Leborgne[1], J. Mille [2] et L. Tougne[3]

[1]INSA de Lyon
[2]Université Lyon 1
[3]Université Lyon 2

Résumé
The skeleton is an essential shape descriptor providing a compact representation of a shape that can be used in real object recognition context. However, due to the discretization, the required properties to use it for graph matching - homotopy to the shape, consequently connectivity, thinness, robustness to noise - may become difficult to obtain simultaneously. In this paper, we propose a new skeletonization algorithm having all this properties, based on the Euclidean distance map. More precisely, the algorithm combines in a clever manner the centers of maximal balls included in the shape and the ridges of the distance map. A post-processing is then applied to thin and prune the resulting skeleton. We compare the proposed method to three fairly recent methods and show its good properties.

Le squelette est un descripteur de formes important qui fournit une représentation compacte de la forme étudiée pouvant être utilisée en reconnaissance d'objets réels. Néanmoins, du fait de la discrétisation, les propriétés requises pour construire un graphe (finesse, robustesse au bruit, homotopie, donc par conséquent connexité) peuvent être difficiles à obtenir simultanément. La squelettisation proposée, basée sur la carte de distance, a toutes ces propriétés. Plus précisément, l'algorithme extrait les centres des boules maximales de la forme ainsi que les crêtes de la carte de distance pour les combiner de manière intelligente. Un post-traitement est utilisé pour amincir et élaguer le squelette. Ces différentes étapes se font en temps linéaire. Le squelette ainsi obtenu a été comparé à d'autres squelettes de la littérature et nous avons mis en évidence ses « bonnes » propriétés pour l'appariement de graphes.

Mots clé : Carte de Distance Euclidienne, boules maximales, squelette, détection de points de crêtes, résistance au bruit.

1. Introduction

Considérons la reconnaissance d'objets 2D obtenus après une étape de segmentation d'images. Une des méthodes pour résoudre ce problème est d'extraire un ensemble de caractéristiques de l'objet à reconnaître (à classifier), et des objets représentatifs de la base de données, puis de les comparer. Ces descripteurs peuvent provenir du contour de la forme (périmètre, élongation, compacité, courbures ...) ou de son intérieur (couleur(s), texture, squelette ...). Contrairement à de nombreux descripteurs, le squelette ne donne pas une information quantitative. Il s'agit d'une compression de la forme permettant de conserver son apparence globale, ses propriétés topologiques et géométriques. En d'autres termes, un algorithme de squelettisation convertit une image binaire contenant un objet en un réseau de lignes décrivant la forme et la topologie de la structure de l'objet [BLL07]. Pour comparer les squelettes de différentes formes, l'idée est de les convertir en graphes (les branches étant les arêtes et, les points de jonction et les points extrémités étant les sommets) qui seront appariés. Cependant, pour qu'un squelette puisse facilement être converti en un graphe, il est nécessaire qu'il ait les propriétés suivantes, qui ne sont pas triviales lorsque la forme est représentée par un ensemble de points dans $\mathbb{Z}^2$:

- connexion : si le squelette n'est pas connecté, le graphe ne sera pas connecté non plus ;
- minceur (branches de 1 pixel d'épaisseur) : un squelette épais génère des problèmes d'extraction de chemins.

De plus, dans le but d'obtenir des appariements efficaces et pertinents dans un contexte d'objets réels, il est nécessaire de construire des squelettes résistants au bruit. En d'autres termes, une légère déformation du bord, ne modifiant pas l'allure générale de la forme, ne doit pas générer de branches. Les algorithmes de la littérature ont beaucoup de difficultés à satisfaire simultanément toutes ces propriétés. Dans $\mathbb{Z}^2$, ils peuvent être classifiés de cette manière :

- méthodes de squelettisation basées sur un amincisse-

ment [LLS92, STRA10] : De manière intuitive, il s'agit de "peler" une forme pour obtenir un ensemble de points connectés d'épaisseur un pixel, qui préserve la topologie de la forme. Autrement dit, l'amincissement est une opération dont le but est de supprimer itérativement tous les points simples non terminaux. Le principal problème de ces méthodes est la sensibilité au bruit.
– méthodes de squelettisation basées sur la carte de distance [CLS03, LLBL07, SBTZ02, CM07, RT05] : L'objectif est d'identifier les points clés sur une carte de distance choisie, dans laquelle chaque pixel est associé à la valeur de sa distance au plus proche pixel de fond. Le problème, ici, est d'arriver à extraire suffisamment de points pour obtenir un squelette connecté et fin.

Comme dit précédemment, la reconnaissance d'objets requiert une représentation de formes qui soit peu sensible aux changements mineurs. Or, l'inconvénient majeur du squelette, de manière générale, est sa sensibilité au bruit sur le bord de la forme. Pour y remédier, il est courant d'utiliser un traitement qui peut être de deux types :
– lissage du bord de la forme : ceci est effectué avant le calcul du squelette pour faire disparaître le bruit du bord non désiré ainsi que les artéfacts de discrétisation [DPS00]. Cependant, dans ce cas, on ne maîtrise pas les effets du lissage sur l'allure générale du squelette.
– suppression des branches non désirées du squelette : il s'agit d'un post-traitement au calcul du squelette (appelé élagage [LWZH13, BLL07]) basé sur des mesures de prépondérances locales ou globales.

L'algorithme proposé, appelé DECS, exploite la carte de distance de deux manières. Tous d'abord, il calcule les centres des boules maximales contenus dans la forme (une boule maximale est un disque contenu dans la forme qui n'est pas entièrement contenu dans une autre boule maximale de la forme). Puis, il connecte les centres de ces boules maximales grâce à une méthode basée sur le filtre LoG appliqué à la carte de distance. Ceci permet d'obtenir un squelette connecté et résistant au bruit.

Avant de détailler la méthode proposée en Section 3, nous décrivons brièvement en Section 2, trois méthodes de la littérature faisant partie de l'état de l'art grâce à leurs propriétés. Ces méthodes seront comparées à la notre en Section 4.

2. Méthodes utilisées pour les comparaisons

Nous avons choisi de comparer notre méthode (DECS) à trois méthodes existantes : K3M est une méthode récente d'amincisssement, les méthodes de Choi *et al* et Hamilton-Jacobi Skeleton sont deux méthodes basées sur la carte de distance, tout comme la méthode proposée.

2.1. K3M [STRA10]

Il s'agit d'une des dernières méthodes d'amincissement. C'est une version modifiée de KMM [SRT01]. L'algorithme commence par détecter les points du bord de la forme puis il les supprime en parallèle si la configuration de leurs voisins le permet. Ces opérations sont répétées jusqu'à ce que l'amincissement soit stable. L'intérêt de cet algorithme est qu'il fournit un amincissment précis. De plus, il est utilisable facilement dans une large gamme d'applications du fait de la clareté et de la simplicité des différentes étapes de calcul.

2.2. Méthode de Choi *et al* : Squelette Euclidien basé sur un critère de connexité [CLS03]

Cette méthode de squelettisation est basée sur une carte de Distance Euclidienne Séquentielle Signée (8SSED) [Ye88]. Un critère de connexité est proposé pour déterminer l'appartenance, ou non, d'un pixel donné au squelette. Ce critère est basé sur un ensemble de paires de points le long du bord de l'objet, qui sont les points de contour les plus proches du pixel considéré et de ses 8 voisins. Cette méthode est intéressante du fait de son utilisation récente dans les travaux de Bai et Latecki concernant l'appariement de graphes [BL08].

2.3. Hamilton-Jacobi Skeleton [SBTZ02]

Cette méthode est basée sur le gradient de la carte de distance Euclidienne. Les vecteurs obtenus sont dirigés vers les crêtes de la carte de distance (*cf* Figure 1). Plus il y a de vecteurs qui convergent vers un point p, plus la valeur de crête de p est élevée. Un point appartient au squelette si sa valeur de crête est assez élevée et qu'il n'est pas un point simple.

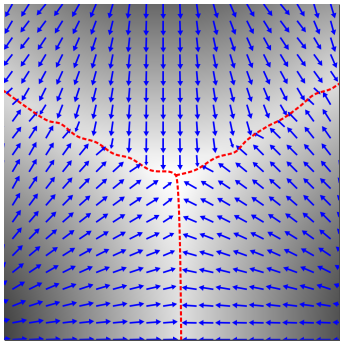


Figure 1: Champ de vecteurs ∇EDT généré grâce au calcul du gradient de la Transformée en Distance Euclidienne. Les vecteurs sont dirigés vers les crêtes de la carte de distance. Les crêtes sont représentées en rouge sur la figure.

3. Méthode permettant l'extraction de DECS

3.1. Vue d'ensemble

Notons $\mathcal{I} \subset \mathbb{Z}^2$ une image de taille $M \times N$ et $\mathcal{S} \subset \mathcal{I}$ une forme 8-connexe (cette hypothèse est utilisée afin d'assurer la connexité du squelette) trouée ou non. Notons p un pixel de $\mathcal{S}$ et $N_8(p)$ l'ensemble des huit voisins 8-connexes de p.

La figure 2 donne une vue générale de la méthode proposée.

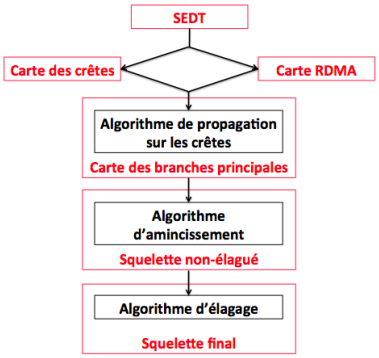


Figure 2: Diagramme de l'algorithme DECS

Figure 3: Représentation de la carte de distance Euclidienne au carré, dans laquelle les valeurs les plus grandes tendent vers le blanc alors que les valeurs les plus petites tendent vers le noir.

Durant la première étape, le calcul de la carte de distance Euclidienne au carré, nommée SEDT, est réalisé [MRH00] (*cf* subsection 3.2). Cette carte est la base des deux étapes suivantes, qui se déroulent en parallèle. D'une part, nous calculons les centres des boules maximales (*cf* subsection 3.3) [CM07] qui sont des points de squelette mais qui ne sont pas connectés. Ces boules sont stockées dans la carte RDMA. D'autre part, nous appliquons un Laplacien de Gaussienne (filtre LoG) sur la carte de distance Euclidienne pour extraire la carte des crêtes (*cf* subsection 3.4). Notons que les crêtes contiennent les centres des boules maximales. Par conséquent, nous utilisons la carte des crêtes pour connecter les centres des boules maximales obtenus auparavant. Nous utilisons un post-traitement, à savoir un amincissement et un élagage, pour obtenir un squelette utilisable pour l'appariement de graphes (*cf* subsection 3.5).

3.2. Transformée en Distance Euclidienne au carré (SEDT)

La distance Euclidienne a été choisie pour sa précision. En effet, elle produit une distance exacte contrairement à la distance de Chanfrein, par exemple. La technique que nous utilisons [CM07, MRH00] est basée sur un algorithme séparable (*i.e.* il procède par passes successives sur chaque dimension). Il a donc une complexité linéaire en n, le nombre de pixels de $\mathcal{S}$. Le but de la SEDT en dimension 2 est de calculer une carte appelée $SEDTmap = \{sedt(i,j)\}_{i,j}$ telle que chaque point (i,j) de $\mathcal{S}$ obtienne une étiquette qui soit la distance Euclidienne au carré au plus proche point appartenant à $\bar{\mathcal{S}}$, où $\bar{\mathcal{S}}$ est le complément de $\mathcal{S}$.

Plus formellement :

$$sedt(i,j) = min\{(i-x)^2 + (j-y)^2; 0 \leqslant x < M, 0 \leqslant y < N \text{ et } (x,y) \in \bar{\mathcal{S}}\}$$

pour chaque point (i,j) de $\mathcal{S}$.
Une représentation est donnée à la Figure 3.

3.3. Axe Médian Discret Réduit (RDMA) [CM07]

Pour chaque point p de $\mathcal{S}$ de coordonnées (i,j), la SEDT en ce point représente le rayon au carré de la plus grande boule centrée en p incluse dans la forme. Comme le squelette contient les centres des boules maximales, l'idée principale est de déterminer ces points en utilisant la $SEDTmap$. Pour cela, il est nécessaire d'avoir un test d'inclusion, qui détermine si une boule est incluse dans une autre, ou pas. Plus formellement, définissons la notion de boule maximale discrète :

Définition 1 (boule maximale discrète)
Soit $d^2 : \mathbb{Z}^2 \times \mathbb{Z}^2 \longrightarrow \mathbb{N}$ une distance Euclidienne au carré discrète.
Une boule de centre c et de rayon r relative à la distance d^2 est définie par :

$$B_2^<(c,r) = \{q \in \mathbb{Z}^2 | d^2(c,q) < r^2\}$$

Une boule maximale discrète est une boule discrète contenue dans la forme, non entièrement recouverte par une autre boule discrète, elle aussi contenue dans la forme.

La Figure 4 illustre ce concept.

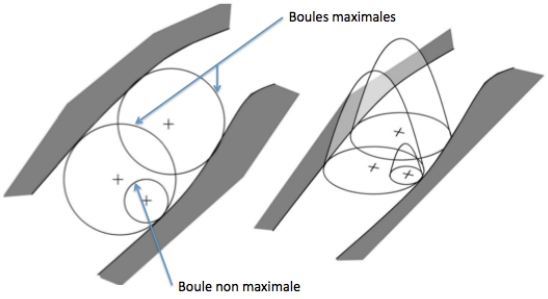


Figure 4: (à gauche) : Illustration du concept de boule maximale. (à droite) : Représentation des boules par des paraboloïdes elliptiques [CM07].

Pour réaliser cette étape, nous avons utilisé la méthode de Coeurjolly [CM07], qui est elle aussi séparable et a une complexité linéaire en n. L'idée générale est de représenter les boules discrètes par des paraboloïdes elliptiques (*cf* Figure

4) pour ne conserver que ceux appartenant à l'enveloppe supérieure. Pour illustrer, si l'on pose un drap, épousant parfaitement les courbes, sur l'ensemble des paraboloïdes, on ne conserve que ceux qui sont en contact avec ce drap. Les centres des boules maximales sont alors les centres des paraboloïdes qui ont été conservés.

Un exemple est présenté à la Figure 5.

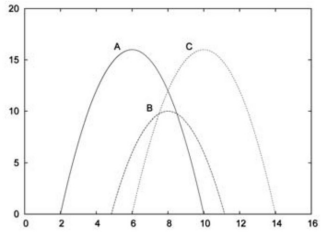


Figure 5: Coupe verticale représentant $\{A,B,C\}$, trois boules maximales. $\{A,C\}$ appartient au squelette, mais pas $\{B\}$ car $\{B\}$ est recouvert par l'union de $\{A\}$ et $\{C\}$ [CM07].

La Figure 6 met en évidence le fait que le RDMA n'est pas connexe, ce qui est son inconvénient majeur. Par la suite, nous extrayons des caractéristiques sur la carte de distance associée à la forme pour permettre la construction d'un squelette connecté.

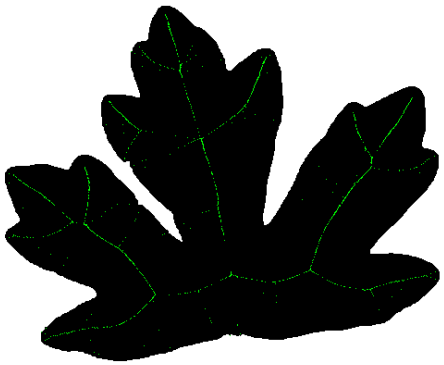


Figure 6: RDMA d'une feuille d'arbre.

3.4. Filtrage Laplacien de Gaussienne (LoG)

Notons que les branches du squelette correspondent aux crêtes de la carte de distance. L'opérateur Laplacien, qui détermine les variations locales de second ordre, permet de les extraire. Il existe différentes manières d'approximer le Laplacien sur une grille discrète. Une des méthodes les plus classiques est l'utilisation de ce masque :

$$\begin{array}{ccc} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{array}$$

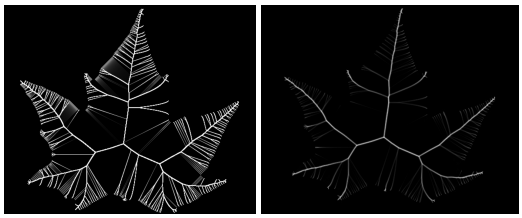
Lorsque nous appliquons ce masque sur la *SEDTmap*, les crêtes sont mises en évidence par des valeurs négatives. En revanche, sur les pentes linéaires, les valeurs sont proches de zéro. Nous pouvons alors considérer le Laplacien négatif comme une mesure de crête. Cependant, le Laplacien est très sensible au bruit quand nous le calculons seul sur la carte de distance, comme nous pouvons le voir à la Figure 7a. Par conséquent, nous utilisons une Gaussienne pour filtrer le bruit. Plus exactement, nous allons convoluer la carte de distance avec le filtre Laplacien de Gaussienne (LoG) négatif d'écart type $\sigma = 1$. Nous obtenons ainsi la carte de crête rdg.

$$\mathrm{rdg}(x,y) = -(EDT * \mathrm{LoG})(x,y)$$

avec

$$\mathrm{LoG}(x,y) = -\frac{1}{\pi\sigma^4}\left(1-\frac{x^2+y^2}{2\sigma^2}\right)\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$$

Un exemple est montré à la Figure 7b. Nous nous apercevons que seules les branches principales sont mises en évidence. Comme le masque obtenu est séparable, la complexité de l'opération de filtrage est linéaire (en $O(\sigma n)$).



(a) (b)

Figure 7: a) représente le Laplacien négatif de la carte de distance Euclidienne et b) est la carte des crêtes, résultant de la convolution de EDT avec un filtre LoG négatif.

À noter qu'un simple seuillage de rdg n'est pas suffisant pour extraire un squelette. En effet, certaines branches principales, qui devraient être conservées, peuvent être déconnectées puisque les valeurs de la carte des crêtes ne sont pas constantes le long des branches. L'idée proposée est donc de combiner le RDMA et la carte des crêtes afin de déterminer la situation des branches principales (où il y a assez de centres de boules maximales et où la carte des crêtes a des valeurs suffisamment élevées) et de connecter l'ensemble en utilisant la carte des crêtes comme guide.

3.5. Algorithme DECS

Dans cette sous-section, nous présentons le coeur de notre algorithme : il s'agit de la propagation des boules maximales sur les crêtes (Algorithme 1). Cet algorithme génère un étiquetage $H = \{h(i,j)\}_{i,j}$ indiquant si un point est un centre de boule maximale ou une crête. Les étiquettes possibles sont $\{NONE, MAX\_BALL, STRONG\_RIDGE, RIDGE\}$, où *MAX_BALL* sont les points notés comme centres de boules maximales, *STRONG_RIDGE* sont les points notés comme ayant une valeur de crête supérieure ou égale à $th_{ridge-high}$, *RIDGE* sont les points notés comme ayant une valeur de crête supérieure ou égale à $th_{ridge-low}$ et inférieure à $th_{ridge-high}$, et *NONE* sont les points notés comme n'appartenant pas au futur squelette. $th_{ridge-low}$ vaut toujours 0.05 alors que $th_{ridge-high}$ varie entre $th_{ridge-low}$ et 1.1. Ces valeurs ont été fixées par expérimentation. L'algorithme utilise une technique de propagation à partir des

centres des boules maximales puisqu'ils appartiennent, de manière sûre, aux branches du squelette. L'algorithme de propagation ne conserve que les points qui sont connectés par un chemin de points de crêtes (*STRONG_RIDGE* ou *RIDGE*) ou de centres de boules maximales (*MAX_BALL*). À la fin de cet algorithme, le squelette est l'ensemble des points ayant une étiquette différente de *NONE*. Il comprend toutes les branches principales mais a un inconvénient. En effet, le squelette a une épaisseur supérieure à 1 pixel. Un exemple est presenté à la Figure 8.

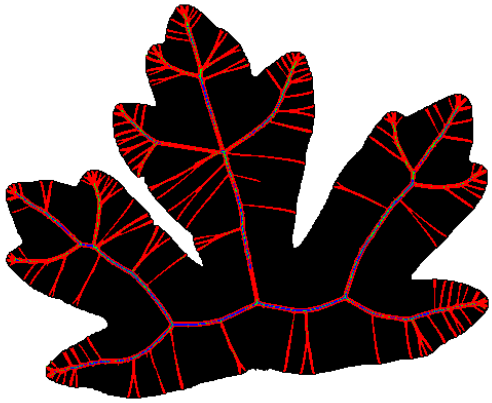


Figure 8: Un exemple de résultat obtenu avec l'Algorithme 1. Les centres des boules maximales apparaissent en vert, les valeurs de crêtes supérieures ou égales à $th_{ridge-high}$ sont visibles en bleu et les valeurs de crêtes comprises entre $th_{ridge-low}$ et $th_{ridge-high}$ en rouge.

Algorithm 1 Pseudocode de l'algorithme de propagation sur les crêtes.

Entrées : RDMA, rdg, $th_{ridge-high}$, $th_{ridge-low}$.
Sorties : $h : \mathcal{S} \longrightarrow \{NONE, MAX\_BALL, STRONG\_RIDGE, RIDGE\}$
Variables : p ,q, $max\_SEDT \in \mathbb{Z}^2$, st une pile de points, $visits$ un ensemble de points.

pour tout $p \in \mathcal{I}$ **faire**
  $h(p) := NONE$
fin pour
$visits := \varnothing$
$max\_SEDT := \underset{p \in RDMA}{\text{argmax}}\ sedt(p)$
$add(st, max\_SEDT)$
$visits := visits \cup \{max\_SEDT\}$
tant que $nonVide(st)$ **faire**
  $p := popTopElement(st)$
  pour tout $q \in N_8(p)$ *tel que* $q \notin visits$ **faire**
    si $q \in RDMA$ **alors**
      $h(q) := MAX\_BALL$
      $add(st, q)$;
    sinon
      si $\text{rdg}(q) \geqslant th_{ridge-high}$ **alors**
        $h(q) := STRONG\_RIDGE$
        $add(st, q)$;
      sinon
        si $th_{ridge-low} \leqslant \text{rdg}(q) < th_{ridge-high}$ **alors**
          $h(q) := RIDGE$
          $add(st, q)$;
        fin si
      fin si
    fin si
    $visits := visits \cup \{q\}$
  fin pour
fin tant que

Pour obtenir un squelette fin, nous avons utilisé l'algorithme d'amincissement MB2 [MMPL99, M.B99]. Cet algorithme permet d'obtenir une courbe dont l'épaisseur est de un ou deux pixel(s) (en fonction de la parité de l'épaisseur de la forme pour la branche correspondante). Il s'agit d'un algorithme itératif parallèle. Il est basé sur la suppression de points simples en fonction de leur configuration (Figure 9). L'algorithme d'amincissement MB2 est détaillé dans le pseudocode de l'Algorithme 2. L'idée principale est de supprimer simultanément les pixels dont les voisins ne s'apparient pas avec les configurations α_1, α_2 tout en conservant la connexité (configuration β), jusqu'à la stabilisation de l'amincissement.

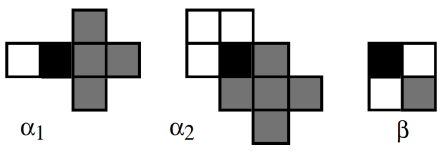


Figure 9: Trois configurations utilisées par l'algorithme d'amincissement MB2. Les pixels noirs et gris appartiennent à $\mathcal{S}$ et les pixels blancs à $\bar{\mathcal{S}}$.

Algorithm 2 Algorithme d'amincissement MB2

Entrées : *Sk* (squelette épais).
Sorties : *Sk* (squelette fin).
Variables : $p_b \in \mathbb{Z}^2$ bord de la forme.

répète
  pour tout p_b de *Sk* **faire**
    $i := 0$
    tant que $i \leq 3$ et p_b n'est pas marqué **faire**
      Faire une rotation de $i \times \frac{\pi}{2}$ des modèles α_1, α_2 *et* β
      si Il y a une coincidence exacte de l'image avec l'intégralité de la configuration α_1 ou α_2 mais pas avec la configuration β **alors**
        Marquer p_b
        $i++$
      fin si
    fin tant que
  fin pour
  Supprimer tous les pixels marqués de *Sk*
tant que l'ensemble des points marqués $\neq$ vide

Le résultat de cet algorithme est un squelette, *Sk*, tel que toutes les branches ont une épaisseur de un pixel en 8-connexité. En d'autres termes, nous ne pouvons plus supprimer de points simples sans modifier la topologie. Un exemple de résultat est montré à la Figure 10. À la fin de cette étape, le squelette est fin et connecté mais peut posséder des branches insignifiantes pour l'appariement de graphes. L'étape suivante, qui est optionnelle, est la suppression de ces branches.

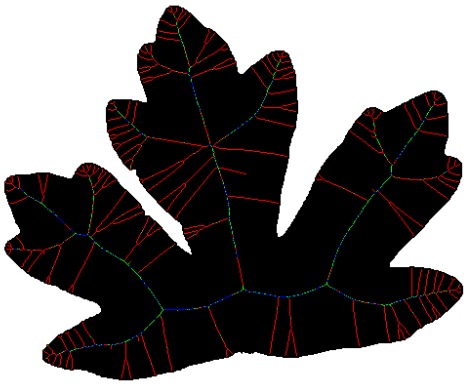


Figure 10: Un exemple de résultat de l'algorithme d'amincissement MB2 appliqué au squelette après propagation sur les crêtes. Les centres des boules maximales apparaissent en vert, les valeurs de crêtes supérieures ou égales à $th_{ridge-high}$ sont visibles en bleu et les valeurs de crêtes comprises entre $th_{ridge-low}$ et $th_{ridge-high}$ en rouge.

L'élagage se réalise en parcourant chaque branche terminale tant que le squelette n'est pas stable. Une branche terminale est supprimée si tous les points ont une valeur de crête inférieure à $th_{ridge-high}$ ou si le pourcentage de boules maximales dans cette branche est inférieur à $th_{perc-max-ball}$. Les seuils peuvent être appris sur un ensemble de données d'entraînement lors du processus de l'appariement de formes. Lors de cette étape d'élagage, chaque point du squelette est visité une seule fois, donc la compléxité est linéaire en $|\ Sk\ |$.

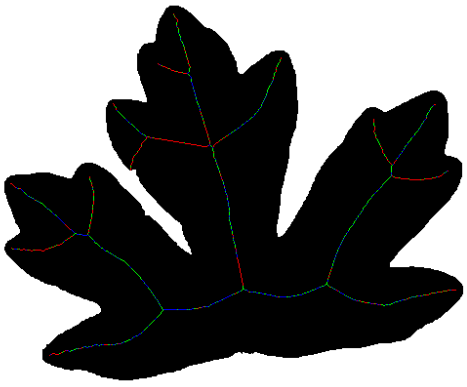
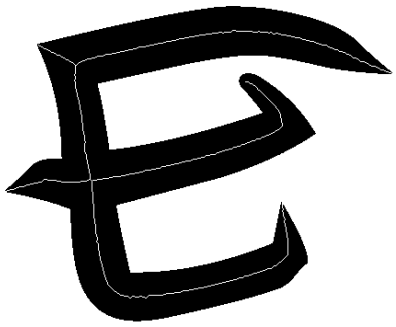
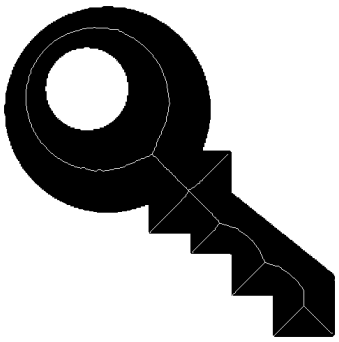


Figure 11: Un exemple de squelette final obtenu grâce à la méthode DECS après élagage. Les centres des boules maximales apparaissent en vert, les valeurs de crêtes supérieures ou égales à $th_{ridge-high}$ sont visibles en bleu et les valeurs de crêtes comprises entre $th_{ridge-low}$ et $th_{ridge-high}$ en rouge.

Finalement, la forme $\mathcal{S}$ a un squelette qui contient tous les points de H qui sont étiquetés *MAX_BALL*, *STRONG_RIDGE* ou *RIDGE*. Un exemple de résultat est montré à la Figure 11. Durant toute cette section, nous avons montré que la complexité totale de la méthode proposée est linéaire en n. Nous pouvons observer, à la figure 12, trois exemples de squelettes obtenus avec la méthode DECS.



(a) $th_{perc-max-ball}$=0,35 ; $th_{ridge-high}$=1,1



(b) $th_{perc-max-ball}$=0,35 ; $th_{ridge-high}$=1

(c) $th_{perc-max-ball}$=0,35 ; $th_{ridge-high}$=1

Figure 12: Exemples de squelettes obtenus avec la méthode DECS. $th_{ridge-low}$ vaut toujours 0.05.

4. Résultats et comparaisons

Dans cette section, nous allons tester differentes propriétés telles que la connexité et la tolérance au bruit. De plus, nous allons comparer notre méthode aux trois méthodes décrites dans la Section 2. Nous allons également faire une comparaison de leur complexité. Tous les tests sont réalisés sur une base de formes créée par Latecki et Lakamper [LLE00] avec, approximativement, un millier de formes.

4.1. Complexité

Nous avons vu que la méthode proposée avait une complexité linéaire en n. En effet, le calcul de la carte de distance ainsi que de la carte des crêtes se déroulent en temps linéaire, tout comme l'extraction du centre des boules maximales. Les algorithmes de propagation, amincissement et élagage parcourent chaque pixel de la forme une fois. Les insertions et suppressions dans les piles et ensembles sont réalisées en temps constant. K3M et la méthode de Choi *et al* ont aussi une complexité linéaire. Cependant, la méthode Hamilton-Jacobi Skeleton a une complexité théorique en O(n log n). Au regard de ce critère, cette dernière méthode est moins intéressante.

4.2. Influence des paramètres

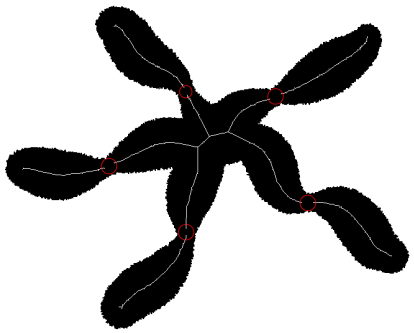
Dans la sous-section 4.4, nous avons effectué des tests de vérité terrain dans lesquels chaque méthode est testée dans ses meilleures conditions. Pour chaque image, nous avons choisi, pour chaque méthode, le(s) seuil(s) qui fourni(ssen)t les meilleurs résultats. Concernant la méthode de Choi *et al*, le nombre de branches inutiles décroît lorsque ρ (seuil de la méthode de Choi *et al*) augmente. Le seuil qui permet d'obtenir le meilleur résultat est $\rho = 803$. L'opposé est vrai pour la méthode Hamilton-Jacobi Skeleton. En d'autres termes, plus th_{AOF} (seuil de la méthode Hamilton-Jacobi Skeleton) décroît, et plus il y a de branches insignifiantes. Les meilleurs résultats pour cette méthode sont obtenus avec un seuil moyen valant -3.06. Concernant la méthode proposée, la suppression des branches inutiles dépend de la combinaison de deux seuils. Ces deux seuils sont complémentaires. Plus $th_{perc-max-ball}$ et $th_{ridge-high}$ augmentent, et moins il y a de branches indésirables. Les seuils moyens retenus pour cette méthode sont $th_{perc-max-ball} = 0.34$ et $th_{ridge-high} = 0.58$.

4.3. Connexité

Les méthodes K3M, Hamilton-Jacobi Skeleton et DECS permettent d'obtenir un squelette connecté dans n'importe quelle situation. En revanche, dans la méthode de Choi *et al*, lorsque le seuil devient grand, la connexité n'est plus garantie. Un élagage idéal permettrait de conserver l'allure générale de la forme tout en supprimant les branches insignifiantes. Avec la méthode de Choi *et al*, lorsque l'élagage est effectué, des déconnexions peuvent apparaître le long du squelette, ce qui altère l'information concernant l'allure générale de la forme. À la vue de ce critère, cette dernière méthode est inutilisable en l'état, pour faire de l'appariement de graphes car elle ne conserve pas la topologie de la forme. Notons que ce critère est primordial pour nous, puisque notre but est de calculer un graphe basé sur le squelette. Cette déconnexion est mise en évidence à la Figure 13.

Méthode	Connexité	image
DECS	oui	
Choi et *al*	non	
Hamilton -Jacobi Skeleton	oui	
K3M	oui	

(a)



(b)

Figure 13: a) Comparaison de la connexité des 4 méthodes étudiées, b) Mise en évidence de la déconnexion du squelette obtenu avec la méthode de Choi *et al* ($\rho = 850$).

4.4. Résistance au bruit

L'avantage d'utiliser une méthode ayant besoin de seuil(s) est que la résistance au bruit est réglable. Cependant, la difficulté est de faire la distinction entre les informations importantes liées au bord de la forme et le bruit.

Pour tester la tolérance au bruit, nous avons créé 15 squelettes théoriques. Pour ce faire, nous avons dessiné manuellement des squelettes 8-connectés ayant des branches d'un pixel d'épaisseur. En utilisant des fonctions linéaires, sinusoïdales et logarithmiques, nous avons attribué à chaque

pixel une valeur de rayon de la boule maximale lui correspondant. Puis, grâce à ces valeurs, nous avons créé les formes "théoriques" en construisant les boules pour chacun des pixels. Le bruit est ajouté de façon aléatoire en bougeant, dans $\mathbb{Z}^2$, chaque pixel de k pixels le long de son vecteur normal. Pour nos tests, nous avons utilisé $noise_1$ où $k \in [-1;1]$ et $noise_2$ où $k \in [-2.5;2.5]$. Nous avons testé les 3 méthodes pouvant s'adapter au bruit : la méthode de Choi *et al*, la méthode Hamilton Jacobi Skeleton et la méthode DECS. Pour chaque méthode, nous avons selectionné un squelette de référence. Pour cela, nous avons fait varier les seuils et nous avons retenu le squelette le plus proche du squelette théorique du point de vue de la Distance de Hausdorff Modifiée (MHD). Une illustration de ces différents squelette est donnée à la Figure 15.

Définition 2 (Distance de Hausdorff Modifiée :MHD) Soit P et Q deux ensembles de points. Nous définissons leur Distance de Hausdorff Modifiée $MHD(P,Q)$ par :

$$MHD(P,Q) = max\{\frac{1}{|P|}\sum_{p\in P} min_{q\in Q}\{d(p,q)\}, \frac{1}{|Q|}\sum_{q\in Q} min_{p\in P}\{d(q,p)\}\}$$

En d'autres termes, la distance de Hausdorff Modifiée est la distance moyenne entre le squelette obtenu et le squelette de référence.

Concernant les squelettes calculés à partir des formes bruitées par $noise_1$ et $noise_2$, nous avons aussi fait varier les seuils et, pour chaque méthode et, pour chaque image, nous avons retenu le squelette le plus proche du squelette de référence en considérant MHD. Une illustration de ces deux squelettes se trouve à la Figure 15. Pour $i \in [1\dots15]$, pour t=$th_{AOF} \in [-6\dots-0,1]$ pour la méthode Hamilton Jacobi Skeleton, pour t=ρ $\in [4\dots3000]$ pour la méthode de Choi *et al* et t étant la combinaison de $th_{perc-max-ball} \in [0\dots1]$ et $th_{ridge-high} \in [0,1\dots1,1]$ pour DECS. Pour chaque bruit : $MHD(bruit, méthode) = moyenne_i(min_t(mhd(i,t,méthode,bruit)))$ où $mhd(i,t,méthode,bruit)$ est la distance de Hausdorff Modifiée entre le squelette de référence et le squelette retenu pour une image i donnée, un seuil t donné, une méthode donnée et un bruit donné. Le résultat de cette expérience est présenté à la Figure 14.

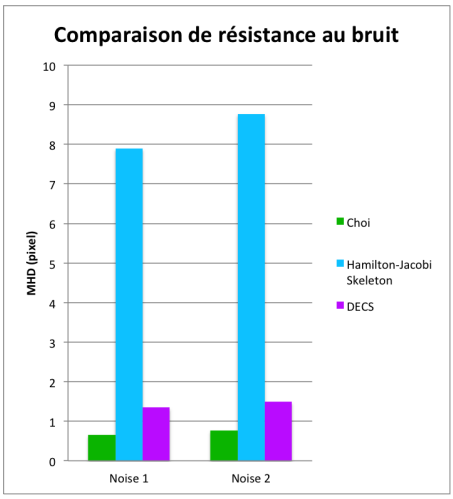
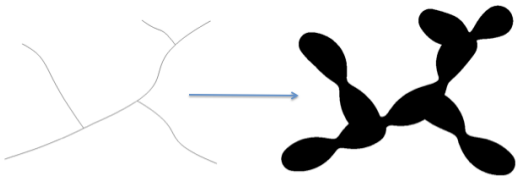
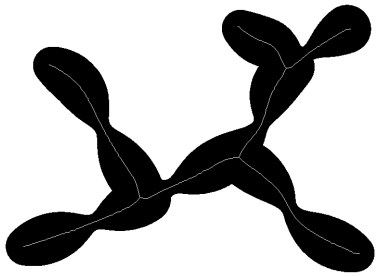


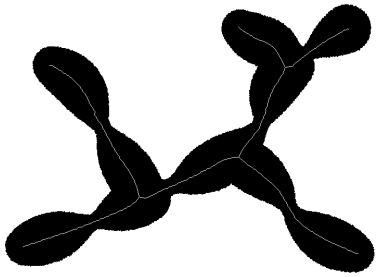
Figure 14: Comparaison en terme de résistance au bruit.



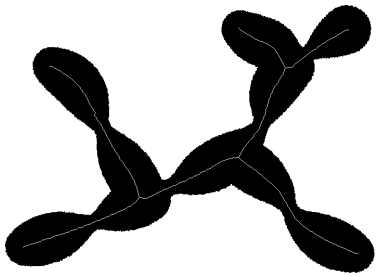
(a)



(b) $th_{perc-max-ball}$=0,2 ; $th_{ridge-high}$=0,4



(c) $th_{perc-max-ball}$=0,3 ; $th_{ridge-high}$=0,5



(d) $th_{perc-max-ball}$=0,5 ; $th_{ridge-high}$=0,7

Figure 15: a) Squelette théorique et construction de l'image théorique associée, b) Squelette de référence, c) Squelette associé au Noise1 et d) Squelette associé au Noise2.

Ce graphique permet d'affirmer que la méthode Hamilton Jacobi Skeleton n'est pas résitante au bruit comparée aux deux autres. De plus, cela permet de remarquer que les méthodes de Choi *et al* et DECS ont un résultat très proche malgré un léger avantage quasiment imperceptible à l'oeil nu pour la méthode de Choi *et al*. Néanmoins, le gros problème de la méthode de Choi *et al* est la non-connexité comme nous l'avons expliqué à la sous-section 4.3. Par conséquent, la méthode de Choi *et al* est inutilisable tel quel dans notre cas.

4.5. Discussion

Les critères du squelette, qui sont prédominants pour l'appariement de formes sont :

- Connexité : la topologie du squelette doit être identique à la topologie de la forme. Les branches doivent correspondre à des parties significatives de la forme et avoir une épaisseur de un pixel. Les discontinuités le long des branches doivent être évitées.
- Faible complexité : l'appariement de formes implique souvent d'analyser un nombre important de formes en un temps limité. C'est pourquoi la complexité linéaire est hautement conseillée.
- Résistance au bruit : qui élimine les branches insignifiantes pour éviter de surcharger le squelette et ainsi, gagner du temps lors de l'appariement.

La méthode DECS présente toutes ces propriétés. Les inconvénients principaux de la méthode Hamilton-Jacobi Skeleton sont sa complexité et surtout son manque de résistance au bruit. Le problème majeur de la méthode de Choi *et al* est la perte de connexité de son squelette lorsque le seuil devient grand, *i.e.* lorsque les branches inutiles ne sont pas prises en compte (ceci est un aspect essentiel pour nous).

5. Conclusion et perspectives

Dans cet article, nous avons présenté un algorithme linéaire permettant d'extraire un Squelette Connecté Euclidien Discret de la forme. Pour réaliser ceci, nous avons proposé un algorithme de propagation sur les crêtes d'une carte de distance Euclidienne et sur les centres des boules maximales. La propagation commence à partir du centre de la boule maximale ayant le rayon le plus grand, ce qui permet de garantir la connexité. Puis, nous obtenons un squelette fin grâce à l'utilisation de l'algorithme d'amincissement MB2. L'étape finale est la réduction du squelette par élagage des branches, basé sur un critère utilisant simultanément les valeurs de crête et les centres des boules maximales. Il est à noter que le squelette proposé se calcule en temps linéaire.

Le squelette proposé a les propriétés désirées (la connexité, la finesse, la résistance au bruit sur le bord de la forme), dans le but d'être utilisé pour l'appariement de graphes. Dans la littérature, il n'y a pas de squelettes ayant toutes ces propriétés, comme nous le mentionnons à la Section 4.

L'algorithme de propagation utilisé pour générer un squelette connecté est uniquement utilisable sur des formes connexes. Pour les formes présentant de multiples composantes, cet algorithme pourrait être amélioré en détectant les composantes connexes présentes dans l'image et en appliquant ainsi DECS sur chacune des composantes. Notre futur travail est de créer un graphe à partir du squelette obtenu. Puis, nous utiliserons cette structure de données pour faire de l'appariement de formes.

Références

[BL08] BAI X., LATECKI L. J. : Path similarity skeleton graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.. Vol. 30*, Num. 7 (2008), 1282–1292.

[BLL07] BAI X., LATECKI L. J., LIU W. : Skeleton pruning by contour partitioning with discrete curve evolution. *IEEE Trans. Pattern Anal. Mach. Intell.. Vol. 29*, Num. 3 (2007).

[CLS03] CHOI W.-P., LAM K.-M., SIU W.-C. : Extraction of the euclidean skeleton based on a connectivity criterion. *Pattern Recognition. Vol. 36*, Num. 3 (2003), 721–729.

[CM07] COEURJOLLY D., MONTANVERT A. : Optimal separable algorithms to compute the reverse euclidean distance transformation and discrete medial axis in arbitrary dimension. *IEEE Trans. Pattern Anal. Mach. Intell.. Vol. 29*, Num. 3 (2007), 437–448.

[DPS00] DIMITROV P., PHILLIPS C., SIDDIQI K. : Robust and efficient skeletal graphs. In *Computer Vision and Pattern Recognition* (2000), pp. 1417–1423.

[LLBL07] LATECKI L. J., LI Q., BAI X., LIU W. : Skeletonization using ssm of the distance transform. In *International Conference on Image Processing (5)* (2007), pp. 349–352.

[LLE00] LATECKI L. J., LAKAMPER R., ECKHARDT U. : Shape descriptors for non-rigid shapes with a single closed contour. In *Computer Vision and Pattern Recognition* (2000), pp. 1424–1429.

[LLS92] LAM L., LEE S.-W., SUEN C. Y. : Thinning methodologies - a comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.. Vol. 14*, Num. 9 (1992), 869–885.

[LWZH13] LIU H., WU Z., ZHANG X., HSU D. F. : A skeleton pruning algorithm based on information fusion. *Pattern Recognition Letters. Vol. 34*, Num. 10 (2013), 1138–1145.

[M.B99] M.BERNARD T. : Improved low complexity fully parallel thinning algorithm. *International Conference on Image Analysis and Processing* (1999), 215–220.

[MMPL99] MANZANERA A., M.BERNARD T., PRÊTEUX F., LONGUET B. : Ultra-fast skeleton based on isotropic fully parallel algorithm. *Proc. of Discrete Geometry for Computer Imagery* (1999).

[MRH00] MEIJSTER A., ROERDINK J., HESSELINK W. : A general algorithm for computing distance transforms in linear time. *Math. Morphology and Its Applications to Image and Signal Processing* (2000), 331–340.

[RT05] REMY E., THIEL E. : Exact medial axis with euclidean distance. *Image and Vision Computing. Vol. 23*, Num. 2 (2005), 167–175.

[SBTZ02] SIDDIQI K., BOUIX S., TANNENBAUM A., ZUCKER S. W. : Hamilton-jacobi skeletons. *International Journal of Computer Vision. Vol. 48*, Num. 3 (2002), 215–231.

[SRT01] SAEED K., RYBNIK M., TABEDZKI M. : Implementation and advanced results on the non-interrupted skeletonization algorithm. *International Conference on Computer Analysis of Images and Patterns* (2001), 601–609.

[STRA10] SAEED K., TABEDZKI M., RYBNIK M., ADAMSKI M. : K3m : A universal algorithm for image skeletonization and a review of thinning techniques. *Applied Mathematics and Computer Science. Vol. 20*, Num. 2 (2010), 317–335.

[Ye88] YE Q. : The signed euclidean distance transform and its applications. *Proceedings of the Ninth International Conference on Pattern Recognition. Vol. 1* (1988), 495–499.

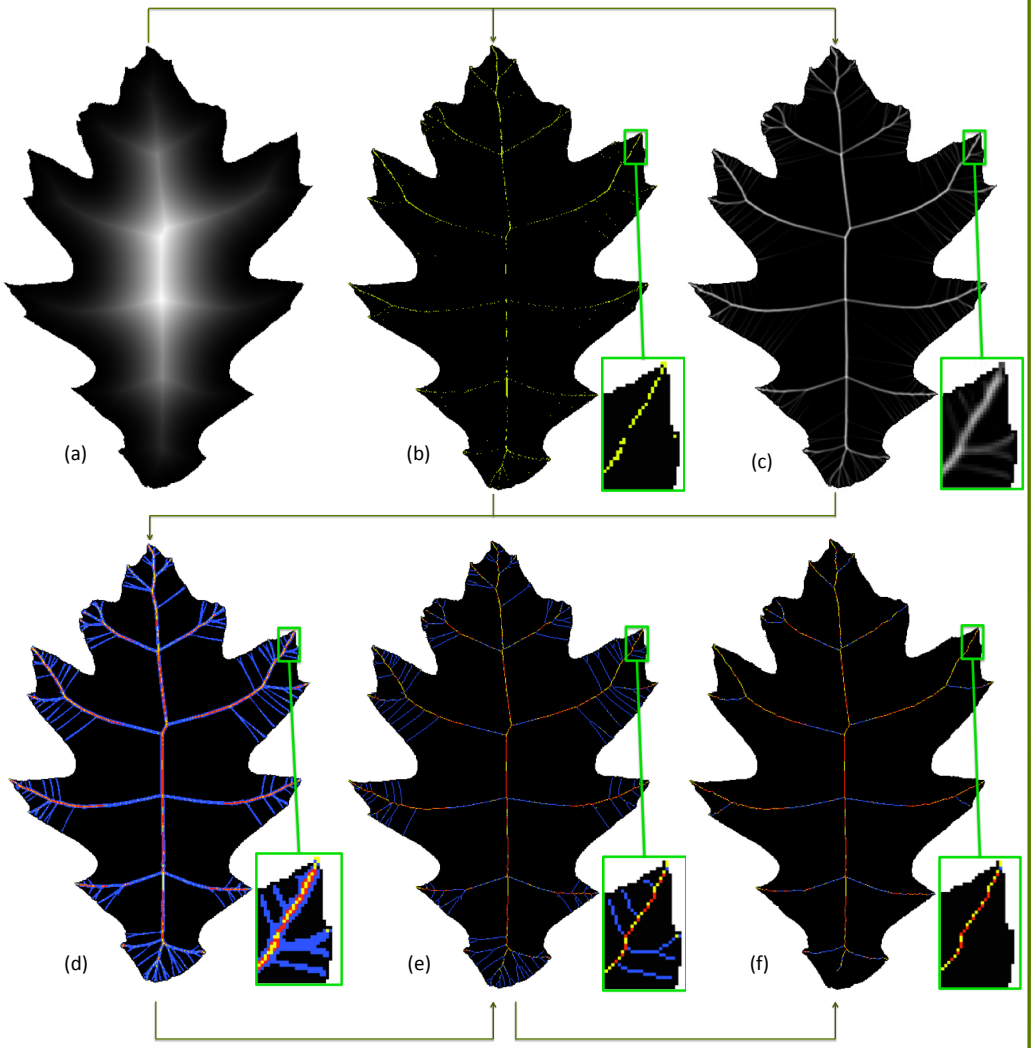
Squelette Euclidien Discret Connecté (DECS) résistant au bruit pour l'appariement de formes basé graphes

Aurélie LEBORGNE
Encadrants: Laure TOUGNE et Julien MILLE
aurelie.leborgne@liris.cnrs.fr - http://liris.cnrs.fr/aurelie.leborgne/

Résumé

Le squelette est un descripteur de formes important qui fournit une représentation compacte de la forme étudiée pouvant être utilisé en reconnaissance d'objets réels. Néanmoins, du fait de la discrétisation, les propriétés requises pour construire un graphe (finesse, robustesse au bruit, homotopie, donc par conséquent connexité) peuvent être difficiles à obtenir simultanément. La squelettisation proposée, basée sur la carte de distance, a toutes ces propriétés. Plus précisément, l'algorithme extrait les centres des boules maximales de la forme ainsi que les crêtes de la carte de distance pour les combiner de manière intelligente. Un post-traitement est utilisé pour amincir et élaguer le squelette. Ces différentes étapes se font en temps linéaire. Le squelette ainsi obtenu a été comparé à d'autres squelettes de la littérature et nous avons mis en évidence ses « bonnes » propriétés pour l'appariement de graphes.

Schéma de l'algorithme



a) Carte de distance Euclidienne au carré, b) Carte des centres des boules maximales, c) Carte des crêtes obtenue après convolution de la carte de distance Euclidienne avec un filtre Laplacien de Gaussienne négatif, d) Résultat obtenu après la propagation des centres des boules maximales sur les crêtes, e) Résultat obtenu après l'amincissement MB2, f) Squelette final après élagage.

Légende de couleur: 

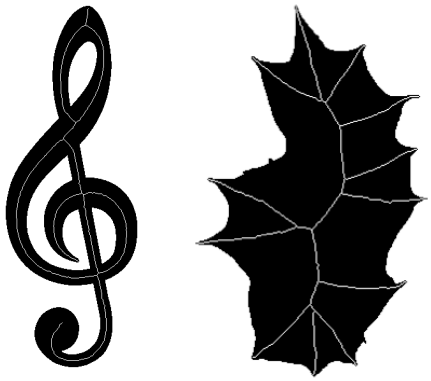
Valeur de crête : centres des boules maximales

Contexte

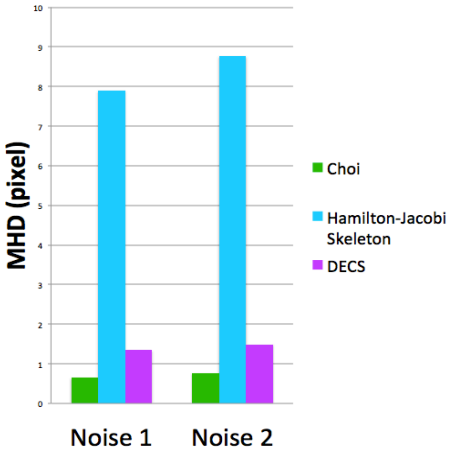
◆ Projet ANR ReVeS – Application Folia (Reconnaissance de végétaux)

Résultats

◆ Exemples



◆ Résistance au bruit



- ☑ Complexité linéaire
- ☑ Homotopie à la forme (par construction)
- ☑ Finesse (par construction)
- ☑ Résistance au bruit

Conclusion et travaux à venir

◆Squelette avec de bonnes propriétés ◆Appariement de graphes

Ce travail est effectué dans le cadre de l'ANR ReVeS (ANR-10-CORD-005) soutenue par l'Agence Nationale de la Recherche.

Comparaison de la segmentation pixel et segmentation objet pour la détection d'objets multiples et variables dans des images

J. Pasquet[1,3], M. Chaumont [1,2], G. Subsol [1]

[1]LIRMM, CNRS/Université Montpellier 2, France
[2]Université de Nîmes, France
[3]Berger Levrault, 31676 Labège, France

Résumé

Cet article étudie et compare deux méthodes de segmentation. La première est la segmentation par objet [VJ01] où l'on cherche à détecter des fenêtres à partir d'un modèle. La seconde est la segmentation pixel [ARTLdM10, FVS08], où l'on cherche à déterminer à quelle classe appartient chaque pixel. De plus, nous proposons une extension au modèle classique de détection par cascade de HOG [ZYCA06] en utilisant les sacs de mots visuels. Des expérimentations sur des jeux de données réelles permettront la comparaison et mettront en avant un gain de performance non négligeable de notre méthode.
In this paper we compare two classical methods of segmentation. First one, the object segmentation, where we try to find an object in a sliding window from a generated model. The second one, the objet level pixel segmentation, where we try to classify all the pixels in an image. Moreover, we propose an extension to the object segmentation, this extension uses bags of features. Finally, we apply some experementations on these two methods using real data set in order to compare the performance.

Mots clé : Segmentation d'images, détection d'objets, sac de mots visuels

1. Introduction

Dans ce papier, nous comparons deux méthodes classiques de segmentation à savoir la segmentation au niveau objet et la segmentation au niveau pixel.

La première approche consiste à construire par apprentissage un modèle des objets recherchés. Pour cela, une base d'apprentissage constituée de multiples imagettes de chaque objet est utilisée. Pour chaque imagette, nous pouvons extraire un vecteur caractéristique décrivant l'objet et servant à l'apprentissage d'un modèle [VJ01, RBK95]. Ensuite, afin d'évaluer notre modèle et de l'utiliser nous cherchons la présence des objets sur une image test. Pour cela, il est nécessaire de lancer un test de présence via une fenêtre glissante de taille variable en tous points de l'image. Le nombre de fenêtres étant très important, pour obtenir un résultat dans des temps raisonnables, nous devons réaliser plusieurs optimisations. De nombreuses solutions existent afin de lutter contre ce problème comme l'utilisation de cascades [VJ01] réduisant fortement la complexité de la phase de tests, ou une présélection des zones d'intérêts dans l'image et donc une diminution des zones à tester [LBH08].

Cependant, une autre solution, la segmentation pixel, consiste à changer l'approche et à classifier directement chaque pixel de l'image [FVS08, RnAK*11]. Ainsi, lors de la face d'apprentissage, un vecteur caractéristique est extrait de chaque pixel et associé à un label. L'ensemble des vecteurs caractéristiques servent à l'apprentissage d'algorithmes. Lors de la face d'évaluation, chaque pixel de l'image test est classifié et associé à une probabilité d'appartenir à un objet. Ainsi, le nombre de vecteurs à tester est faible en comparaison à l'approche objet. Toutefois, la phase d'évaluation retourne une carte de probabilité qu'il est nécessaire de post-traiter afin d'obtenir des régions.

Afin, de comparer les deux méthodes nous utilisons une base de données constituée d'images aériennes couleur haute résolution de cimetière. Les objets recherchés étant les tombes, il a été montré dans [CTS*13] que ce problème est difficile. En effet, les tombes sont des objets à forte variabilité, multiples et très proches.

Dans une première section 2, nous présenterons la segmentation pixel ainsi que les descripteurs utilisés. Puis, dans la section 3 nous présenterons en détail l'approche par

construction d'un modèle. Nous proposons également une extension du modèle objet dans la partie 3.2. Dans notre approche, nous proposons d'introduire les sacs de mots visuels multi-fenêtre pour décrire les objets. Ensuite, nous présenterons notre base d'apprentissage ainsi que les résultats dans la section 4. Finalement, nous conclurons et proposerons des perspectives.

2. Approche pixel

L'approche pixel consiste à classifier directement chaque pixel de l'image [FVS08, RnAK*11]. Le schéma classique de fonctionnement de cette approche se divise en quatre étapes : l'extraction d'un vecteur caractéristique par pixel, la quantification des vecteurs en mots visuels, la statistique des mots visuels présents dans le voisinage du pixel étudié, c'est à dire la création d'histogrammes représentant la fréquence d'apparition des mots visuels, et la classification des histogrammes de mots visuels.

2.1. Extraction des caractéristiques

De nombreuses caractéristiques peuvent définir l'information un pixel. Dans son article [ARTLdM10], Aldavert et coll. montrent que les histogrammes de gradient orienté (HOG) sont bien adaptés. De plus, le temps de calcul de cette caractéristique extraite, de façon dense, peut être considérablement amoindri grâce à l'utilisation d'image intégrale [ZYCA06]. Les vecteurs HOG permettent de décrire la forme locale d'une zone à partir d'une distribution d'orientation et d'amplitude de gradient. Pour extraire le vecteur HOG d'une région (ou bloc) de dimension D_H nous divisons cette région en cellules de dimension réduite D_C. Dans chacune des cellules, nous formons un histogramme des N orientations du gradient (équation 1). Pour $N = 8$, chaque pixel vote pour son orientation principale, c'est à dire 0°, 45°, ... ou 325°.

$$\theta = floor\left(\frac{arctan(\frac{G_y}{G_x}) + \pi}{2\pi}.N\right).\frac{2\pi}{N} \quad (1)$$

Avec G_x et G_y les gradients selon l'axe x et y et $floor$ la fonction qui arrondit à l'entier inférieur.

Le vote de chaque pixel est pondéré par la norme du gradient (équation 2) afin de différencier les fortes et faibles discontinuités. Les histogrammes de chaque cellule sont ensuite concaténés puis normalisés avec la norme L2. Cet histogramme normalisé forme le vecteur HOG. La figure 1 récapitule le processus de formation du vecteur HOG.

$$G = \sqrt{G_x^2 + G_y^2} \quad (2)$$

Avec G_x et G_y les gradients selon l'axe x et y.

2.2. Création du dictionnaire

Chaque pixel est caractérisé par un vecteur HOG. Afin d'utiliser l'approche par sac de mots visuels, nous devons trouver les meilleurs représentants des vecteurs HOG. Pour obtenir les K meilleurs représentants nous utilisons une méthode de classification non supervisée tel que le

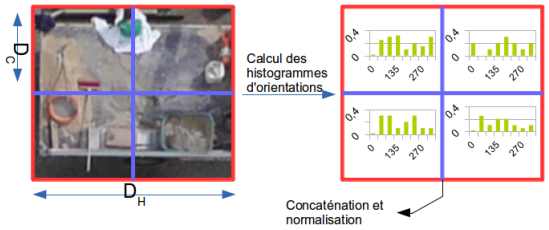


Figure 1: *Les contours rouges représentent la zone où le HOG est extrait et les contours bleus les délimitations dans le cas où nous souhaitons 4 cellules.*

classique Kmeans. Cependant, la complexité du Kmeans étant importante pour un nombre élevé de clusters (représentants) [MM09], nous pouvons utiliser des méthodes de Kmeans hiérarchiques [NS06] ou des forêts d'arbres aléatoires [MNJ08] afin d'obtenir avec des coûts raisonnables un grand nombre de mots visuels. Dans notre cas d'étude nous utiliserons des ERF.

2.3. Création du vecteur caractéristique et apprentissage

Une fois le dictionnaire créé, chaque vecteur HOG de chaque pixel est quantifié en un mot visuel. Afin d'effectuer l'apprentissage, il est nécessaire d'obtenir une liste de vecteurs caractéristiques pour chaque objet. Pour ce faire, nous passons une fenêtre glissante de largeur L_f sur chaque pixel de l'image. Nous créons dans la fenêtre, un histogramme des fréquences d'apparition des mots visuels. Cet histogramme est le vecteur caractéristique du pixel central à l'image. Lors de la création de la base d'apprentissage, connaissant la vérité terrain, nous associons au vecteur caractéristique le label du pixel central. Une fois la base d'apprentissage construite nous effectuons l'apprentissage avec un classifieur linéaire. Le schéma 2 récapitule tout le processus d'extraction de vecteurs caractéristiques dans le cas d'une approche pixel.

2.4. Phase de test

Lors de la phase d'évaluation, à l'aide du dictionnaire de la phase d'apprentissage nous quantifions l'image de test en mots visuels. Une fenêtre glissante sert ensuite à évaluer chaque pixel de l'image en créant un histogramme de fréquence. Cet histogramme de fréquence constitue le vecteur caractéristique qui est classifié par le classifieur. Notons que Aldavert et coll. [ARTLdM10] proposent une version optimisée de la phase d'évaluation. En effet, grâce à l'utilisation d'un classifieur linéaire à chaque mot visuel peut correspondre un score constant. En remplaçant chaque mot visuel par son score nous pouvons construire une image intégrale [VJ01] de l'image de scores. L'évaluation d'une région autour de pixel correspond à la somme des valeurs des scores des mots visuels de la région. Elle ne prend alors que quatre calculs et s'effectue en temps réel. De la phase de tests, il

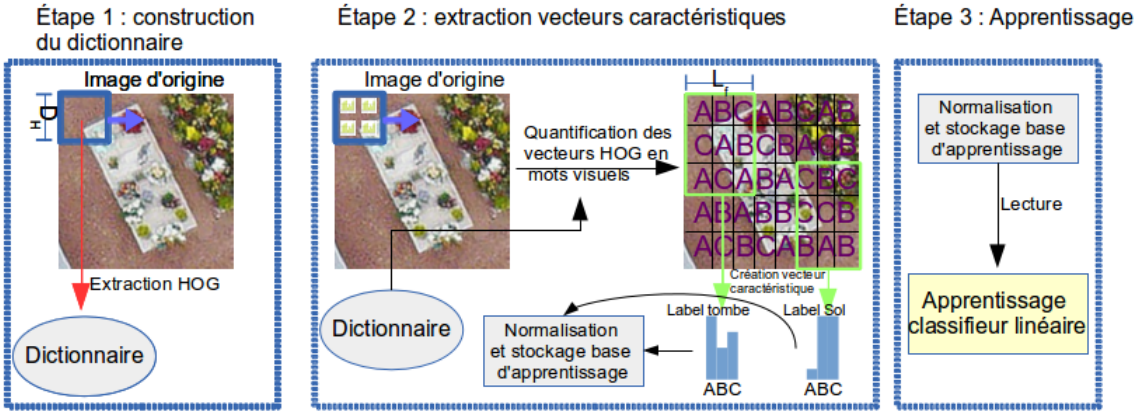


Figure 2: *Récapitulatif du processus d'apprentissage afin d'effectuer une segmentation par pixel*

résulte une carte de probabilité, il faudra effectuer un post-traitement afin de détecter et localiser les objets.

3. Segmentation objet

Dans l'approche objet ou modèle, le modèle de l'objet est construit. Contrairement à l'approche pixel, où chaque pixel de l'objet est caractérisé par un vecteur, nous allons former un seul vecteur caractéristique pour chaque fenêtre englobant un objet recherché. Notre base d'apprentissage sera constituée d'imagettes des objets recherchés ainsi que d'imagettes n'étant pas pas des objets recherchés. Dans la section 3.1, nous présentons le protocole classique d'extraction de vecteurs caractéristiques par objet. Dans la section 3.2, nous proposons une amélioration de cette méthode en y rajoutant une approche par sac de mots.

3.1. Extraction des caractéristiques

Sur une image, un objet peut avoir des tailles différentes de part sa nature où à cause d'effets de zoom. Afin d'être robuste à ce changement d'échelle, les objets sont redimensionnés à une taille constante. Ensuite, comme pour l'approche pixel, il est nécessaire d'extraire des caractéristiques comme les histogrammes de gradients orientés pour décrire nos objets. Afin d'obtenir une description plus fine de chaque objet, nous utiliserons la méthode de Qiang Zhu et coll. [ZYCA06] pour extraire de multiples HOG. Dans leur papier [ZYCA06], Qiang Zhu et coll. extraient des HOG de tailles différentes d_i avec $i \in 0..T$ et T le nombre de tailles possibles dans toutes les positions de la fenêtre. Une fois extraits tous les HOG sont concaténés en un vecteur de très grande dimension. La figure 3 résume ce processus. Afin de limiter le temps nécessaire lors de la phase de test, Qiang Zhu utilise un système de cascades de classifieurs [VJ01]. Ce processus d'extraction de caractéristiques est identique à celui mis en place par Viola et Jones [VJ01] avec un descripteur plus HOG robuste. Ceci créé un modèle multi-fenêtre robuste de nos objets.

Une fois extraits, les vecteurs caractéristiques ne sont pas quantifiés en mots visuels et servent à l'apprentissage d'un classifieur linéaire.

3.2. Extension de la segmentation objet

Nous proposons une extension du modèle de Qiang Zhu et coll. [ZYCA06] en y rajoutant la statistique et distribution de mots visuels. Notre idée est de décrire l'objet à l'aide de mots visuels de différentes résolutions. Pour chaque taille d_i de vecteur HOG extrait nous construisons en amont un dictionnaire de N mots visuels. Lors de la phase d'extraction de vecteurs caractéristiques, nous remplaçons chaque vecteur HOG par sa quantification en mots visuels. Pour chacun des T dictionnaires (pour chaque résolution) nous construisons un histogramme des fréquences d'apparition des mots visuels. Tous les histogrammes sont ensuite concaténés pour former le vecteur caractéristique de dimension $N * T$ (voir schéma 8).

Le fait d'utiliser comme vecteur caractéristique une statistique, provoque une perte de l'information sur la localité des vecteurs HOG. Plusieurs solutions existent pour lutter contre ce phénomène []. Dans notre cas, nous avons choisi de représenter les mots visuels dans une pyramide [LSP06] ce qui agrandit considérablement la dimension des vecteurs caractéristiques mais permet de garder partiellement la localité des vecteurs (voir schéma 5).

4. Implémentation et résultats

Dans cette section, nous allons décrire la base de données que nous avons utilisée ainsi que la façon dont nous avons évalué les différentes approches. Ensuite, nous donnerons et discuterons de nos résultats.

4.1. La base de données et protocole d'évaluation

Dans notre cas d'étude, nous allons chercher à détecter les tombes dans des images de cimetière de dimension 4000^2 pixels. Un exemple de la base de donnée est montré en 6. La détection de tombes, comme montré dans [CTS*13], est une

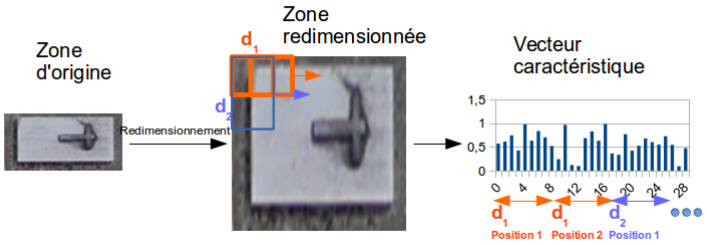


Figure 3: *Récapitulatif du processus d'extraction de multiples vecteurs HOG dans une région.*

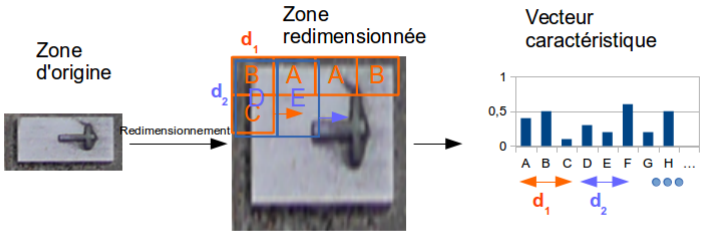


Figure 4: *Processus d'extraction de multiples vecteurs HOG et leur quantification en mots visuels par un dictionnaire dépendant de leur taille.*

Figure 6: *Image provenant de [CTS*13], montrant le cimetière de Lecey.*

tâche difficile. En effet, les tombes sont de tailles différentes (hauteur allant de 80-140 pixels et largeur de 50-60 pixels), avec des ombres portées et souvent des contours abîmés par le temps ou recouverts de terre les rendant quasiment identiques au sol. De plus, l'espace entre les tombes varie de 2 à 20 pixels nécessitant une détection précise. Pour cela, nous utilisons des méthodes d'apprentissage automatique. La société Berger-Levrault a fourni des images aériennes haute définition (2.5cm/pixel) en couleurs (rouge, vert, bleu) de 21 cimetières du département de Haute-Marne. Le nombre de tombes dans la totalité des images avoisine les 1300. Afin de ne pas biaiser nos résultats, les bases d'apprentissage et de test sont distinctes. Ainsi, 19 images serviront à l'apprentissage et 2 images (soit environ 120 tombes) serviront aux tests.

4.2. Paramètres utilisés

Dans cette section, nous décrivons les paramètres utilisés pour chacune des approches. Dans chacun des trois cas, à savoir l'approche pixel, objet et notre proposition d'amélioration, le classifieur utilisé est un SVM linéaire de *liblinear* [FCH*08].

4.2.1. Approche pixel

Dans l'approche pixel, la taille des blocs d'extraction des HOG est fixée à 32 pixels, avec pour chaque HOG une concaténation de quatre cellules. Chaque HOG est ensuite quantifié avec un dictionnaire de type forêt ERF [MNJ08] de profondeur maximale seize avec dix arbres. Le dictionnaire est donc formé de 655360 mots visuels. Le vecteur caractéristique est construit en représentant la fréquence d'apparition de ces mots visuels dans une fenêtre de dimension 40.

4.2.2. Approche objet

Chaque tombe est redimensionnée en fenêtre de dimension 70x70 pixels. Dans cette fenêtre, nous faisons varier des

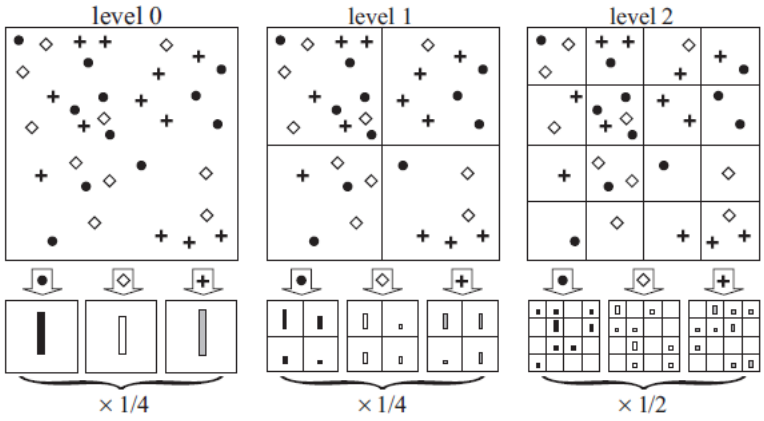


Figure 5: *Schéma provenant de [LSP06], montrant la formation d'un histogramme de mots visuels et leurs poids à partir d'une pyramide de profondeur 3.*

tailles de dimension minimale (5,5) jusqu'à (80,80) afin de tester 1260 tailles différentes. Chaque HOG constitué d'une seule cellule est extrait avec une surface de recouvrement de 50%. Au final, nous obtenons un vecteur caractéristique de dimension 430000.

4.2.3. Proposition d'extension de l'approche objet

Les paramètres utilisés sont les mêmes que ceux de l'approche par modèle de la section 4.2.2 avec un dictionnaire. Le dictionnaire choisi est un Kmeans Hierarchique [NS06]. Celui-ci permet d'obtenir un grand nombre de mots visuels avec un faible coût en complexité. Nos dictionnaires de profondeur 3, ont à chaque niveau 7 centroïdes, soit un total de 343 mots visuels par dictionnaire. La taille du vecteur caractéristique finale est donc de l'ordre de 430000.

4.3. Résultats

Nous allons comparer les différentes approches. En premier, nous comparerons l'approche pixel et l'approche objet. En second, nous comparerons l'approche objet à l'amélioration que nous proposons.

4.3.1. Comparaison approche pixel et objet

Afin de comparer l'approche pixel et l'approche objet il est nécessaire de segmenter la carte de probabilité de l'approche pixel pour obtenir des régions. Pour cela, en utilisant l'image d'origine et la transformée de Hough nous avons cherché à détecter des rectangles dans l'image. A l'aide de la carte de probabilité, nous associons à chaque rectangle un score et gardons uniquement les meilleurs protagonistes. Un exemple des résultats obtenus par les deux approches est donné dans la figure 7. Une fois la carte de probabilité segmentée, nous comparons l'approche de Qiang Zhu et coll. et l'approche pixel dans le tableau ci-dessous.

	Segmentation pixel	Segmentation objet
Précision (%)	0.23	0.3
Rappel (%)	0.41	0.41

Pour un rappel fixé, nous constatons que la segmentation objet est plus précise. Une explication possible est causée par la proximité des tombes. En effet, l'espace entre deux tombes voisines est compris entre 2 et 20 pixels. Aussi, l'approche pixel ne distingue pas les frontières entre les objets laissant cette tâche à la post segmentation. L'approche objet, de par son fonctionnement intrinsèque, connaît le modèle d'une tombe et donc les détecte directement. Dans le cas d'une détection d'objets multiples très proches, l'approche pixel n'est, à priori, pas très adaptée et demande un post traitement fort.

4.3.2. Comparaison approche multiple HOG avec et sans sac de mots

Lors d'une approche objet, le classifieur peut à chaque zone donner une probabilité d'être une tombe ou autre chose. La probabilité seuil minimal est requise pour considérer qu'une zone est une tombe. Dans la figure 8, nous faisons varier ce paramètre de seuil et traçons les courbes ROC de l'approche de Qiang Zhu et coll. (en bleu) et de notre approche décrite en 3.2 (en rouge). Pour un rappel fixé notre approche est en moyenne 9% plus précise que l'approche de Qiang Zhu et coll. Ce résultat est significatif et représente un gain relatif en précision moyen de 17%.

Sur la figure 9, on constate que la taille des dictionnaires influence directement sur les performances de notre méthode. Aussi, un faible nombre de mots visuels provoque une chûte de performance. Ce résultat est classique, en effet, une faible taille de dictionnaire ne permet pas de décrire correctement un système. Cependant dans le cas d'un dictionnaire trop grand, les performances diminuent également. Une explication possible serait que le vecteur caractéristique devienne trop grand et que l'information ne serait pas assez dense. Cependant mieux étudier ce phénomène semble être intéressant pour l'avenir.

5. Conclusion

Dans cet article, nous avons comparé l'approche pixel et l'approche modèle. Dans notre cas d'étude, il résulte

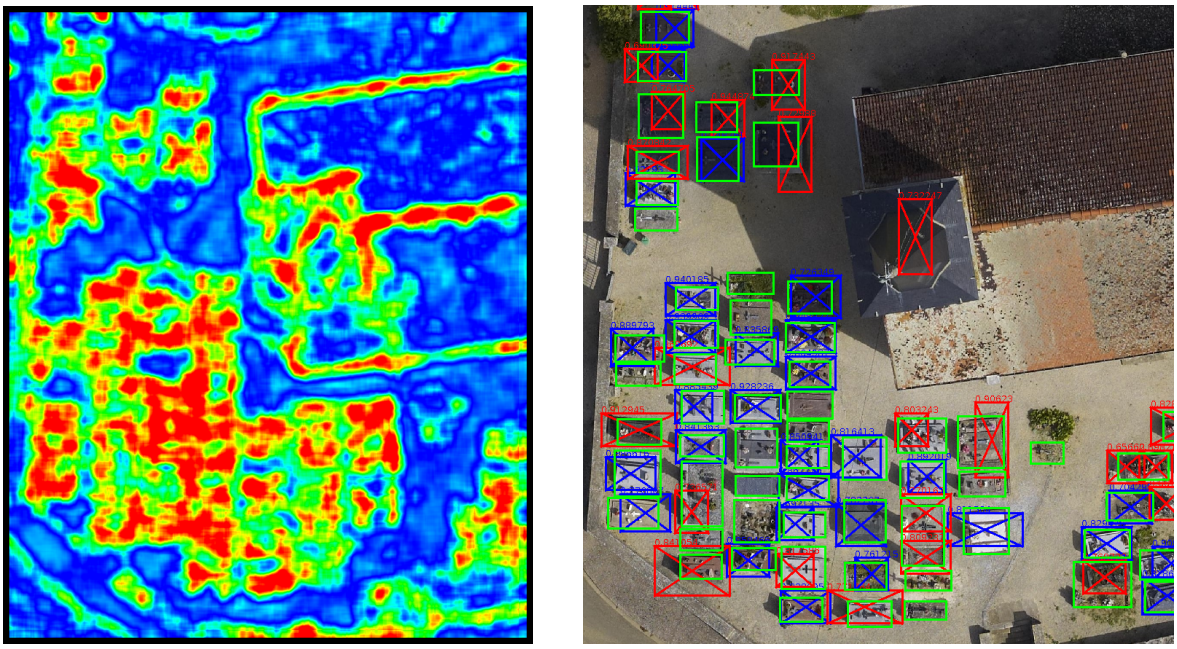


Figure 7: *Les deux figures illustrent l'approche pixel à gauche et l'approche objet à droite. L'approche pixel génère une carte de probabilité où les pixels rouges sont ceux avec une forte probabilité d'être une tombe et inversement les pixels bleus ont une forte probabilité de ne pas être une tombe. L'approche objet, génère directement des régions, les rectangles bleus sont les tombes détectées, les rectangles rouges sont les faux positifs et en vert nous affichons la vérité terrain.*

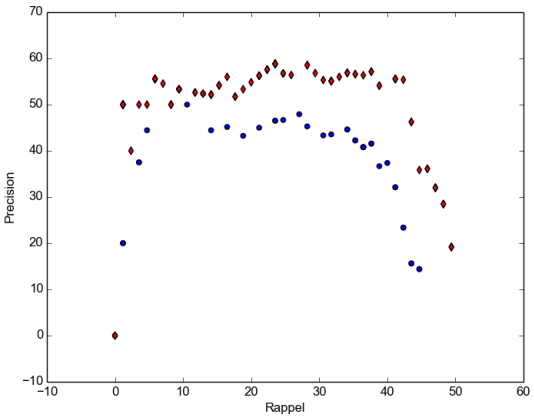


Figure 8: *La courbe ROC de l'approche décrite en 3 est représentée par les points ronds bleus, la courbe ROC de notre approche décrite en 3.2 est représentée par les diamants en rouges.*

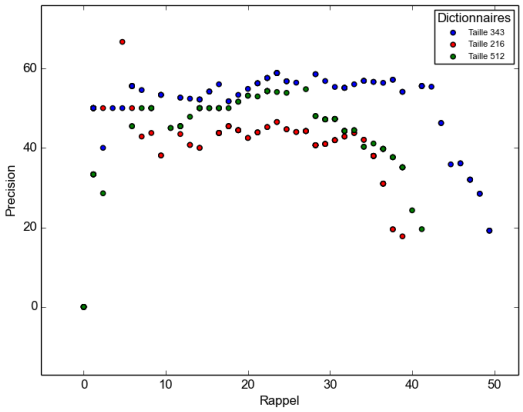


Figure 9: *Représente les courbes ROC de notre proposition avec des dictionnaires de type Kmeans Hiérarchique de différentes tailles.*

que pour le cas d'objets très proches et multiples, l'approche pixel est moins adaptée et précise de 7 %. Nous nous sommes donc focalisés sur la segmentation par modèle et proposons une extension du modèle classique de cascades de HOG. Dans la partie 4.3.2, nous montrons que l'utilisation de mots visuels augmente de façon significative, soit un gain relatif en précision de 17%. Afin de mieux comprendre ce gain, il serait intéressant, d'étudier la transformée inverse des représentants(du dictionnaire) des vecteurs HOG en image [VKMT13]. Afin d'améliorer notre extension du modèle de cascades de HOG, plusieurs pistes sont possibles. La première consiste à changer la méthode de construction du dictionnaire ainsi que sa taille ; laquelle influence grandement, comme nous l'avons montré dans la partie 4.3.2. De plus, la taille du dictionnaire peut varier en fonction de la taille des fenêtres glissantes.

6. Remerciements

Nous souhaitons remercier tout particulièrement la société Berger Levrault pour son soutient dans nos travaux. De plus, nous remercions également Laurent Deruelle, Francois Bibonne et Pol Kennel pour leurs conseils.

Références

[ARTLdM10] ALDAVERT D., RAMISA A., TOLEDO R., LÓPEZ DE MÁNTARAS R. : Fast and Robust Object Segmentation with the Integral Linear Classifier. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2010), pp. 1046–1053.

[CTS*13] CHAUMONT M., TRIBOUILLARD L., SUBSOL G., COURTADE F., PASQUET J., DERRAS M. : Automatic localization of tombs in aerial imagery : Application to the digital archiving of cemetery heritage. In *Digital Heritage International Congress (DigitalHeritage), 2013* (Oct 2013), vol. 1, pp. 657–660.

[FCH*08] FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R., LIN C.-J. : LIBLINEAR : A library for large linear classification. *Journal of Machine Learning Research. Vol. 9* (2008), 1871–1874.

[FVS08] FULKERSON B., VEDALDI A., SOATTO S. : Localizing objects with smart dictionaries. In *Proceedings of the 10th European Conference on Computer Vision : Part I* (Berlin, Heidelberg, 2008), ECCV '08, Springer-Verlag, pp. 179–192.

[LBH08] LAMPERT C. H., BLASCHKO M., HOFMANN T. : Beyond sliding windows : Object localization by efficient subwindow search. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (June 2008), pp. 1–8.

[LSP06] LAZEBNIK S., SCHMID C., PONCE J. : Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, pp. 2169–2178.

[MM09] MEENA MAHAJAN PRAJAKTA NIMBHORKAR K. V. : The planar k-means problem is np-hard. *Lecture Notes in Computer Science* (2009).

[MNJ08] MOOSMANN F., NOWAK E., JURIE F. : Randomized clustering forests for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.. Vol. 30*, Num. 9 (septembre 2008), 1632–1646.

[NS06] NISTER D., STEWENIUS H. : Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (Washington, DC, USA, 2006), CVPR '06, IEEE Computer Society, pp. 2161–2168.

[RBK95] ROWLEY H. A., BALUJA S., KANADE T. : Human face detection in visual scenes. In *NIPS* (1995), Touretzky D. S., Mozer M., Hasselmo M. E., (Eds.), MIT Press, pp. 875–881.

[RnAK*11] RUSIÑOL M., ALDAVERT D., KARATZAS D., TOLEDO R., LLADÓS J. : Interactive Trademark Image Retrieval by Fusing Semantic and Visual Content. In *Advances in Information Retrieval*, Clough P., Foley C., Gurrin C., Jones G., Kraaij W., Lee H., Mudoch V., (Eds.), vol. 6611 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2011, pp. 314–325.

[VJ01] VIOLA P., JONES M. : Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (2001), vol. 1, pp. I–511–I–518 vol.1.

[VKMT13] VONDRICK C., KHOSLA A., MALISIEWICZ T., TORRALBA A. : HOGgles : Visualizing Object Detection Features. *ICCV* (2013).

[ZYCA06] ZHU Q., YEH M.-C., CHENG K.-T., AVIDAN S. : Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (Washington, DC, USA, 2006), CVPR '06, IEEE Computer Society, pp. 1491–1498.

INTRA RESIDUAL PREDICTION IN HEVC BY MODE DEPENDENT TEMPLATE MATCHING AND MODE DEPENDENT VECTOR QUANTIZATION

B. Huang[1,2], C. Guillemot[1], F. Henry[2], P. Salembier[3] et G. Clare[2]

[1]INRIA
[2]Orange Labs
[3]Universitat Politecnica de Catalunya

Résumé
La Norme de codage vidéo HEVC définit au total de 35 *modes de prédiction intra image qui visent à réduire la redondance spatiale en exploitant la corrélation dans les pixels à partir des références voisinages. Toutefois, dans cet article, nous montrons tout d'abord qu'il reste des corrélations non-locales après la prédiction intra image, conduisant à des résidus en haute énergie. Nous proposons ensuite deux schémas pour réduire la redondance dans le domaine residuel : l'intra prédiction résiduelle et le codage de résidu par la quantification vectorielle. Les résultats expérimentaux montrent que la prédiction de second ordre atteint* 0,2% *de réduction de débit par rapport au modèle de test HEVC* 8.0*. Le codage de résidu par la quantification vectorielle améliore la performance de* 1.1% *en moyenne.*
The High Efficiency Video Coding standard (HEVC) supports a total of 35 Intra prediction modes which aim at reducing spatial redundancy by exploiting pixel correlation within a local neighborhood. However, in this paper, we first show that non-local correlation remains in the residual signals of intra prediction, leading to some high energy prediction residues. We then propose two strategies to reduce the remaining redundancy in the residual prediction domain : Intra Residual Prediction (IRP) by template matching, and residual coding using Mode Dependent Vector Quantization (MDVQ). Experimental results show that the proposed intra residual prediction achieves 0.2% *bitrate reduction at high bit rate compared with the HEVC Test Model* 8.0*. And the MDVQ-based residual coding improves the performance to* 1.1% *bitrate reduction on average.*

Mots clé : HEVC, intra prediction, residual coding, template matching, vector quantization

1. Introduction

The High Efficiency Video Coding (HEVC) standard [SOHW12] developed by the Joint Collaborative Team on Video Coding has been ratified as an international Video Coding Standard in 2013. Compared to the previous H.264/AVC standard, HEVC improves the coding efficiency of Intra prediction by introducing a larger number of prediction modes used together with a quad-tree based flexible block partitioning structure. The quad-tree based partitioning allows the splitting of a Large Coding Unit (LCU) into smaller Coding Units (CU) to represent complex structures. Up to 35 intra prediction modes are supported in HEVC to produce a more accurate prediction for smooth regions as well as directional structures. The conventional intra prediction in HEVC is efficient at reducing the local spatial redundancy in the original pixel domain. However, the accuracy of intra prediction is limited in regions having complex textures or structures. So the residues obtained by intra prediction in these regions have larger magnitudes than residues in homogeneous areas.

Different methods have been proposed in the literature for improving intra prediction, essentially for H.264/AVC. These methods can be divided into two main categories. The methods in the first category, such as Template Matching [LXWS10] and Epitome-based image compression [CGT*11], improve the coding efficiency by exploiting the non-local similarity within an image, whereas the methods in the second category, such as Second Order of Prediction (SOP) [CWLY08] [KHS09], try to reduce the correlation between the residual signals of inter prediction and the surrounding reference pixels.

In this paper, we focus on the remaining redundancy in the residual prediction domain, which we try to remove by exploiting local and non-local correlation in residual signals. Two strategies are proposed : the Intra Residual Prediction (IRP) by template matching, and the residual coding using Mode Dependent Vector Quantization (MDVQ). The propo-

sed method of intra residual prediction exploits similarities between residual blocks resulting from intra prediction in HEVC by reusing reconstructed residual blocks. Essentially, a subset of previously decoded residue blocks is used to predict the current original residue block. Template matching helps identifying the most promising candidates within the subset. The proposed method of MDVQ-based residual coding provides another way of reducing the remaining redundancy in residual signals. Mode-dependent codebooks are learned from a training set of residue vectors which are extracted from training video sequences. These codebooks are optimized in a rate-distortion sense and do not need to be adapted for each Quantization Parameter (QP).

The rest of the paper is organized as follows. Section 2 recalls the main principles of HEVC intra prediction and quad-tree partitioning. The proposed intra residual prediction is described in Section 3. Experimental results of intra residual prediction are presented in Section 4. The approach of MDVQ-based residual coding and its performance are briefly discribed in Section 5. We finally conclude this paper and discuss potential directions for future research in Section 6.

2. INTRA PREDICTION AND QUAD-TREE PARTITIONING

In HEVC, an intra predicted CU can have one or four PUs, each of which specifies a region with an individual intra prediction mode. For each PU, the best intra prediction mode is selected under a mechanism of Most Probable Mode [LBH*12]. The CU is further split into a quad-tree of Transform Units (TU), on which transform, scalar quantization and entropy coding are performed. Table 1 represents the supported CU size in HEVC and the corresponding TU size inside a CU [CGT*11].

CU size	TU size
64×64	32×32, 16×16, 8×8
32×32	32×32, 16×16, 8×8
16×16	16×16, 8×8, 4×4
8×8	8×8, 4×4

Table 1: *Supported CU size in HEVC, and available TU size inside a CU*

As mentioned in the introduction, there are remaining redundancies in the residual prediction domain. It is interesting to know the number of bits per residual-pixel required on average to code a TU. Statistical results for sequences in class D are shown in Table 2.

QP	TU 4×4	TU 8×8	TU 16×16	TU 32×32
22	1.51	0.79	0.48	1.21
27	1.07	0.47	0.31	0.54
32	0.74	0.17	0.09	0.05
37	0.43	0.17	0.09	0.05

Table 2: *Number of bits used per residual signal for a TU block*

We noticed that larger TU blocks need fewer bits per pixel than 4×4 TU blocks, so they are more efficient in the sense of bit cost. That means a 4×4 TU block has more remaining redundancy than larger TU blocks after intra prediction.

3. INTRA RESIDUAL PREDICTION BY MODE DEPENDENT TEMPLATE MATCHING

The proposed intra residual prediction by mode-dependent template matching further reduce spatial redundancy within a picture by exploiting self-similarities between residual signals of intra prediction. The scheme of intra residual prediction is shown in Figure 1. An image block is first intra predicted, which can be refered to as First Order Prediction (FOP), and the first order residue of intra prediction $\mathbf{r}_f$ is computed. The proposed intra residual prediction, which can be seen as a Second Order Prediction (SOP) is then performed on the first order residual signals. The SOP reuses a decoded first order residual block in the causal area as a predictor. The best candidate in the causal area is selected by applying a Template Matching algorithm, which is described in Section 3.3. The prediction error of the SOP, named as second order residue $\mathbf{r}_s$, is pocessed by the usual series of operations in HEVC : transform, scalar quantization and entropy coding.

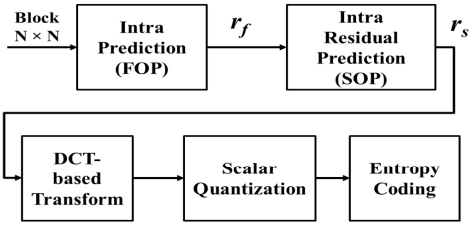


Figure 1: *Scheme of Intra Residual Prediction*

3.1. Predictor Candidate Search Region

In our method, the intra residual prediction is performed solely for 4×4 TU blocks. We reuse the reconstructed residual blocks which are aligned on a 4×4 grid as the predictor candidates. As shown in Figure 2, the search region of prediction candidate covers a causal area of the current original residual block. The size of the search region is defined by a width of $M \times 4$ and a height of $N \times 4$. Thus, a total of $M \times N$ predictor candidates are available in this search region.

3.2. Shape of Template for a Residual Block

As shown in Figure 2, the template of the original residual block to be predicted refers to the surrounding residues adjacent to the current block. Using the same principle, we define the corresponding residues in the candidate's neighborhood as the predictor template. An effective template could represent the texture or structure of the corresponding block. As a consequence, the shape of the template should adapt to the intra prediction mode of the residual block. Intra prediction modes of DC and Planar are frequently used for regions

that have homogeneous texture. As shown in Figure 3(a), we use the L-formed template for these two modes. Horizontal and vertical intra prediction modes use reference pixels from the left or from above the current block respectively, and here we apply the templates in Figure 3(b) and 3(c). For diagonal modes, the shape of template depends on the directional structures of the current block, such as (d) (e) in Figure 3.

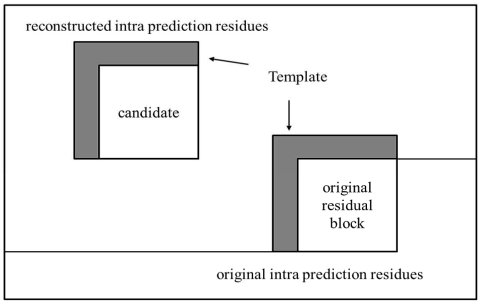


Figure 2: *Shape of template for residue block and candidate*

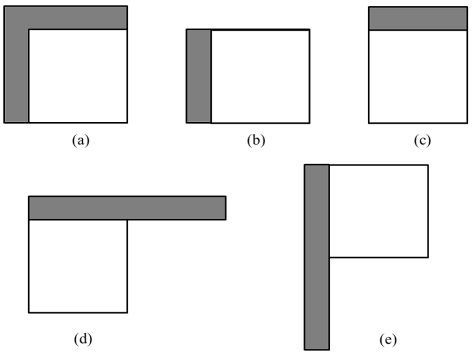


Figure 3: *Shape of mode dependent template*

3.3. Algorithm of Template Matching and Candidate List Construction

Let us denote $\mathbf{r}_f$ the original first order residual signals in the current block to be predicted by intra residual prediction. Its template containing reconstructed first order residues is denoted by $\mathbf{t}_f$. Let $\mathbf{r}_c(x,y)$ represents the reconstructed first order residues in the candidate block, which have the left upper corner residual at the position (x, y). The reconstructed first order residues in the template of this candidate is denoted as $\mathbf{t}_c(x,y)$. Then, the best candidate c^* for block $\mathbf{r}_f$ is selected from all available candidates in the search region by applying the following steps.

In step 1, the validity of a candidate is checked by using Eq.(1), which measures a square sum of all residues in the template of candidate. Candidates which do not satisfy this condition will be rejected.

$$||\mathbf{t}_c(x,y)||^2 \neq 0 \tag{1}$$

In step 2, for all available candidates, the template matching is performed by computing the Sum of Squared Error (SSE) between the residual signals in the template of the current block $\mathbf{t}_f$ and those in the template of a candidate $\mathbf{t}_c(x,y)$. Then all candidates are sorted by SSE in an ascending order.

$$SSE(\mathbf{t}_f, \mathbf{t}_c(x,y)) = ||\mathbf{t}_f - \mathbf{t}_c(x,y)||^2 \tag{2}$$

In step 3, a candidate list $\mathcal{L}$ is constructed which contains only the first N candidates in terms of lowest SSE computed by Eq.(2). A candidate is identified by its index in this list.

In step 4, the best candidate c^* in the candidate list $\mathcal{L}$ is the one containing reconstructed residues $\mathbf{r}_{c^*}$ that is most similar to the original residual singals $\mathbf{r}_f$ in the current block to be predicted. This can be expressed as :

$$c^* = \arg\min_{c\in\mathcal{L}} ||\mathbf{r}_f - \mathbf{r}_c(x,y)||^2 \tag{3}$$

The difference between the reconsturcted first order residues in the best candidate residual block and the original first order residues in the current block is refered to as second order residue $\mathbf{r}_s$. It should be noted that, the intra residual prediction (SOP) is performed only if it reduces the energy of the first order residue. Otherwise, the intra residual prediction is skipped. This condition can be expressed as :

$$||\mathbf{r}_f - \mathbf{r}_{c^*}(x,y)||^2 < ||\mathbf{r}_f||^2 \tag{4}$$

3.4. Structure of Signaling and Decoding Scheme

By analyzing the position of the 4×4 TU blocks in which the intra residual prediction is activated, we found that for some CUs, the nested TUs tend to have the same decision on the use of intra residual prediction. Consequently, we use three syntax elements and a tree structre for the signalization of the proposed intra residual prediction. For a CU block containing 4×4 TU blocks, the syntax element irp_cu_flag indicates whether one of the nested TUs on which the IRP is performed. This flag is coded using one bit and an associated context-adaptive binary arithmetic coding (CABAC) context. When irp_cu_flag is set to 1, for each 4×4 TU block inside this CU, a syntax element irp_tu_flag is signalled to indicate whether the IRP is performed. This syntax element is coded using one bit and an associated CABAC context. When the IRP is used, a fixed-length syntax element irp_tu_idx is transmitted. It represents the index of the best predictor candidate in the candidate list.

At the decoder side, the syntax element irp_cu_flag of a CU block is parsed firstly. If the flag value is zero, none of the inside TUs on which the intra residual prediction is performed. Otherwise, the syntax element irp_tu_flag is then parsed for each 4×4 TU block. Finally, the index of the best candidate is parsed for TU block having syntax element

irp_tu_flag equals 1. A candidate list is constructed using the procedure described in Section3.3. The decoded residues of the current block actually correspond to the second order residues of the intra residual prediction. By adding the predictor identified by the decoded index to this second order residue, the first order residue is reconstructed.

4. EXPERIMENTAL RESULTS OF INTRA RESIDUAL PREDICTION

The proposed intra residual prediction was implemented in the reference software HM8.0 [C7] of the HEVC test Model. Since our algorithm aims at improving the intra coding efficiency, all the frames of a sequence were intra coded. The encoder is configured following the JCT-VC Common Test set with the All-Intra profile. The performance is evaluated by comparing our algorithm with HM8.0. Eight video sequences (ParkScence, Cactus, BasketballDrill, BQMall, BasketballPass, BlowingBubbles, BQSquare, RaceHorses) with different resolutions as specified in HEVC [F.B11] were encoded. The experiments are performed under two ranges of quantization parameters : mid-bitrate (MBR) for QP values 22, 27, 32 and 37 ; and high-bitrate (HBR) for QP values 16, 21, 26, 31. The BD-Rate performances of the proposed algorithm are measured with the method in [Bjø01]. Table 3 shows the performance where the negative values mean a bitrate saving.

Sequences	**HBR** (%)	**MBG** (%)
ParkScene	−0.06	−0.02
Cactus	−0.07	−0.04
BasketballDrill	−0.18	−0.12
BQMall	−0.11	−0.06
BasketballPass	−0.09	−0.04
BlowingBubbles	−0.08	−0.04
BQSquare	−0.07	−0.04
RaceHorse	−0.07	−0.01

Table 3: *Simulation results in iterms of bitrate saving*

From the results, we can see that the proposed algorithm provides a bit rate saving of up to 0.18% under high-bitrate compared with HM8.0. It should be noted that, without the cost of signalization in the intra residual prediction, a bitrate reduction of 13% could be achieved, a theoretical upper limit that gives an indication about the possible improvements. We conclude that the template matching is not the most efficient method for identifying the best candidate of intra residual prediction, due to higher cost of signalling the predictor.

5. MDVQ-BASED RESIDUAL CODING

The approach of residual coding using MDVQ provides another way of exploiting the correlations in residual signals. Actually, in a block of samples intra predicted with a planar mode or DC mode, the residual signals have a relatively homogeneous structure, whereas those derived from angular prediction modes tend to have directional structures. In our approach, the codebooks of vector quantization are learned with the aim of modelling the directional characteristics of the intra prediction residual signals.

The scheme of MDVQ-based residual coding is similar to the intra residual prediction shown in Figure 1, whereas the IRP is replaced by the MDVQ-based residual coding. For a image block which is intra predicted with mode i, the first order original residue of intra prediction $\mathbf{r}_f$ is then quantized by a matching codevector $\mathbf{v}$ in a pre-generated codebook C_i. The vector quantization error (the difference between the original residue and the matching codevector) which can be refered to as the second order residue $\mathbf{r}_s$, is processed by the following operations : transform, scalar quantization and entropy coding. In our experiments, the MDVQ-based residual coding has been used for the TU block size of 4×4 and 8×8. For each of them, 35 codebooks corresponding to the 35 intra prediction modes are derived. Each codebook contains 256 codevectors, which represents a good compromise between storage requirements, complexity and performance. The codebook learning processe is iterated, based on a training set of original residue vectors. They are obtained by extracting from training video sequences those residue vectors selected by the Lagrange Rate Distortion Optimization.

The experiments are performed under the mid-bitrate range of quantization parameter. The encoder is configured as the experiments of intra redisual prediction described is Section 4. Table 4 shows the performance of MDVQ-based residual coding of sequences of six test classes. Here the MDVQ codebooks are trained on sequences that are different from those used to measure the performance. This is the realistic use case. One can observe an average bitrate saving of 1.1%. Interestingly, a larger gain is observed for low-resolution sequences which are usually more difficult to compress, this being attributed to the larger proportion of 4×4 and 8×8 TU blocks. The method also performs well on videoconference (class E) and screen content (class F) sequences.

By utilizing codebooks which are well suited to the test sequences, such as the training sequence for codebook generation is the same as the test sequences, better performance can be achieved. Table 5 shows the experimental results of MDVQ-based residual coding on class B sequences. One can observe that our method provides a bitrate reduction of 4.9% on average in this ideal-but not realistic-case. This experiments demonstrate that there are still rooms for improving the MDVQ-based residual coding.

6. CONCLUSION

The proposed intra residual prediction scheme aims to exploit non-local correlation in the residual domain within a picture. By making use of the adequate shape of template in template matching, intra residual prediction can further reduce the redundancy of residual signal of the conventional prediction approaches. The MDVQ-based residual coding provides a better scheme to reduce the remaining redundancy in residual signals, while further tests indicate that codebook adaptivity could substantially improve the performance.

sequence class	bit-rate saving (%)
NebutaFestival	−0.09
PeopleOnStreet	−1.62
SteamLocomotiveTrain	0.44
Traffic	−1.51
average of class A	−0.7
BasketballDrive	0.12
BQTerrace	−1.33
Cactus	−1.03
Kimono	0.02
ParkScene	−1.44
average of class B	−0.5
BasketballDrill	−2.27
BQMall	−1.34
ParkScene	−0.95
RaceHorse	−0.52
average of class C	−1.3
BasketballPass	−1.26
BlowingBubbles	−1.24
BQSquare	−0.98
RaceHorse	−1.68
average of class D	−1.3
FourPeople	−1.61
Johnny	−1.67
KristenAndSara	−1.2
average of class E	−1.6
BasketballDrillText	−2.53
ChinaSpeed	−1.29
SlideEditing	−0.91
SlideShow	−1.80
average of class F	−1.6

Table 4: *Bitrate savings using sequence-independent codebooks*

sequence name	bit-rate saving (%)
Kimono1_1920x1080p	−0.25
ParkScene_1920x1080p	−3.19
Cactus_1920x1080p	−13.66
BQTerrace_1920x1080p	−1.99
BasketballDrive_1920x1080p	−5.16

Table 5: *Bitrate savings using sequence-dependent codebooks*

Références

[Bjø01] BJØNTEGAARD G. : Calculation of average psnr differences between rd-curves. *ITU-T SG16 Q.6 Doc. VCEG-M33* (avril 2001).

[C7] Hevc test model (online). `http://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware`.

[CGT*11] CHERIGUI S., GUILLEMOT C., THOREAU D., GUILLOTEL P., PEREZ P. : Epitome-based image compression using translational sub-pel mapping. *MMSP* (octobre 2011).

[CWLY08] CHEN S., WANG J., LI S., YU L. : Second order prediction (sop) in p slice. *ITU-T SG16/Q.6 Doc. VCEG-AI27* (juillet 2008).

[F.B11] F.BOSSEN : Common test conditions and software reference configurations. *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T VCEG and ISO/IEC MPEG. Doc. JCTVC-G1200* (février 2011).

[KHS09] KIM S., HWANG S., SUNWOO M. : Novel residual prediction scheme for hybrid video coding. *ICIP* (novembre 2009).

[LBH*12] LAINEMA J., BOSSEN F., HAN W., MIN J., UGUR K. : Intra coding of the hevc standard. *IEEE Transaction on Circuits and Systems for Video Technology. Vol. 22*, Num. 12 (décembre 2012).

[LXWS10] LAN C., XU J., WU F., SHI G. : Intra frame coding with template matching prediction and adaptive transform. *ICIP* (septembre 2010).

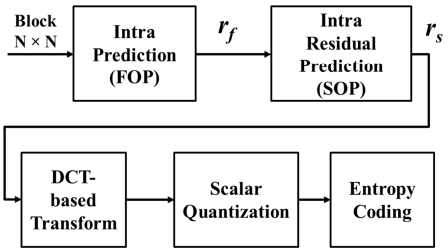
[SOHW12] SULLIVAN G., OHM J., HAN W., WIEGAND T. : Overview of the high efficiency video coding (hevc) standard. *IEEE Transaction on Circuits and Systems for Video Technology* (décembre 2012).

Codage de résidu dans HEVC par Mode Dependent Template Matching et Quantification Vectorielle

B. Huang, F. Henry, C. Guillemot, P. Salembier, C. Gordon

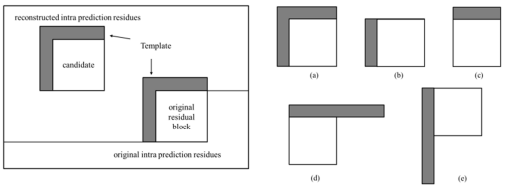
La norme de codage vidéo HEVC définit au total 35 modes de prédiction intra image et supporte le partitionnement d'image avec une grande flexibilité. Ceci permet de réduire la redondance spatiale en exploitant la corrélation dans le domaine pixel, mais des corrélations demeurent, que nous exploitons à travers deux approches pour augmenter le taux de compression.

Codage de résidu par Mode Dependent Template Matching



❖Construction de liste de candidats par Template Matching

$$SSE(\mathbf{t}_f, \mathbf{t}_c(x,y)) = ||\mathbf{t}_f - \mathbf{t}_c(x,y)||^2$$



❖Intra prédiction résiduelle

$$c^* = \arg\min_{c \in \mathcal{L}} ||\mathbf{r}_f - \mathbf{r}_c(x,y)||^2$$

$$||\mathbf{r}_f - \mathbf{r}_{c^*}(x,y)||^2 < ||\mathbf{r}_f||^2$$

❖**Performance**: 0.2% de réduction de débit par rapport à HEVC

Codage de résidu par Mode Dependent Quantification Vectorielle

❖Codage de résidu par Quantification Vectorielle

$$\mathcal{D}(\mathbf{s}_i, \mathbf{v}) = ||\mathbf{s}_i - \mathbf{v} - \mathcal{E}'(\mathbf{s}_i, \mathbf{v})||^2$$
$$= ||\mathcal{E}(\mathbf{s}_i, \mathbf{v}) - \mathcal{E}'(\mathbf{s}_i, \mathbf{v})||^2$$

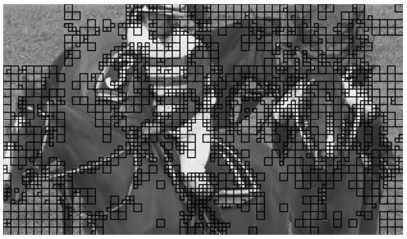
$$\mathbf{v}^* = \arg\min_{\mathbf{v} \in \mathcal{C}_i} \mathcal{D}(\mathbf{s}_i, \mathbf{v}) + \lambda \cdot \mathcal{R}(\mathbf{s}_i, \mathbf{v})$$

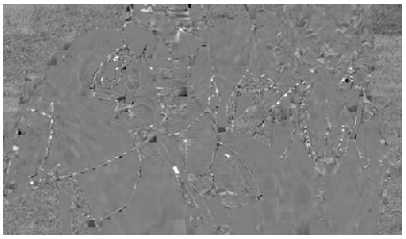
$$\mathcal{D}(\mathbf{s}_i, \mathbf{v}^*) + \lambda \cdot \mathcal{R}(\mathbf{s}_i, \mathbf{v}^*) < \mathcal{D}(\mathbf{s}_i, \mathbf{s}_i') + \lambda \cdot \mathcal{R}(\mathbf{s}_i)$$

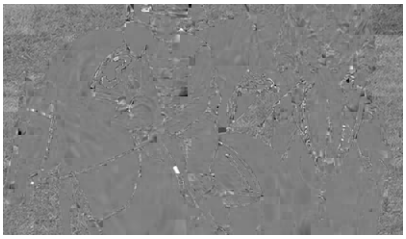
❖**Performance**: 1.1% de réduction de débit par rapport à HEVC

❖Construction de dictionnaires

- Optimisé au sens de débit distorsion
  - Construction itérative
  - Séquence d'apprentissage = résidus sélectionnés par RDO
- Dictionnaires indépendant du paramètre de quantification (QP)







Conclusions et perspectives

❖Le codage de résidu par la quantification vectorielle est plus efficace par rapport à la méthode de Template Matching pour réduire les corrélations dans le domaine résiduel.

❖La performance de quantification vectorielle peut être amélioré si les dictionnaires sont adaptés au séquences.

contact: bihong.huang@orange.com

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche RENNES - BRETAGNE ATLANTIQUE

UMR IRISA

orange™

Vers une reconnaissance d'état affectif à base de mouvements du haut du corps et du visage

Benjamin Allaert, Ioan Marius Bilasco, Adel Lablack

Laboratoire d'Informatique Fondamentale de Lille, Université de Lille 1, France

Résumé

L'émotion est une réaction complexe qui engage à la fois le corps et l'esprit. Elle peut être définie comme étant une réaction affective transitoire d'assez grande intensité provoquée par une stimulation venue de l'environnement. L'analyse des expressions corporelles a un rôle important dans le processus de reconnaissance de l'état affectif. Pour cela, nous proposons une approche de reconnaissance émotionnelle combinant deux canaux : le visage et le corps. Notre contribution s'appuie sur l'analyse du mouvement au sein du visage et du haut du corps qui sont synthétisés par des modèles de direction et de magnitude construits à partir des flux optiques. Ces modèles permettent de s'abstraire des bruits de détection à l'aide de l'extraction des caractéristiques principales des mouvements et constituent une base stable pour identifier les évolutions de l'état émotionnel et plus particulièrement de la valence et de l'arousal. Les modalités sont analysées individuellement et sont ensuite fusionnées dans un deuxième temps afin d'étudier l'apport informationnel issu de l'étude du mouvement de la personne dans sa globalité. L'approche proposée a enfin été validée avec succès sur un sous-ensemble de la base de données SEMAINE et permet de vérifier l'apport informationnel issu du mouvement dans la reconnaissance d'états affectifs.

The emotion is a complex reaction that involves both the body and the spirit. It can be defined as an affective reaction of high intensity usually caused by a stimulus that comes from the environment. The analysis of the body expression through movements is important in an affect recognition process. We propose an approach for affect recognition from two channels : face and body. Our contribution uses an analysis of facial and body movements through direction and magnitude models of motion constructed from optical flows. These models allow to remove the noise using the extraction of the main motion features and constitute a stable base to identify the evolutions of the emotional state and more specifically the valence and the arousal dimensions. Each modality is analyzed alone then combined to study the informative contribution of the user motion. The proposed approach has been validated successfully on a subset of SEMAINE database.

Mots clé : Reconnaissance d'émotions, analyse gestuelle, analyse du mouvement, analyse du visage

1. Introduction

Actuellement, il y a un vrai engouement autour des technologies permettant de reconnaître les émotions à partir de flux vidéo. La reconnaissance des expressions faciales et corporelles joue un rôle important dans une variété d'applications telles que l'automatisation de la recherche comportementale, le traitement audio-visuel de la parole, la téléconférence, l'e-learning, la sécurité aéroportuaire et le contrôle d'accès, etc.

La reconnaissance d'émotions à partir de flux vidéo s'est basée essentiellement sur les expressions faciales durant ces deux dernières décennies [HEF02] [KCY00] mais les systèmes proposés manquent de robustesse dans des environnements non contrôlés et en présence d'émotions spontanées, de variations de poses et d'éclairage, en présence de personnes âgées, etc. Bien qu'une étude fondamentale faite par Ambady et Rosenthal [AR92] ait suggérée que les expressions du visage et les gestes du corps semblent être les plus pertinents pour analyser le comportement humain, la reconnaissance d'émotions via les mouvements du corps ne commence que récemment à attirer l'attention des chercheurs [VDHdG11] [GP05] [CVC07].

Dans notre étude nous ciblons principalement les domaines de l'e-learning et de téléconférence. Dans ces domaines, il est important de garder l'attention de son auditoire pendant toute une présentation, ou du moins, d'être informé de la manière dont le public perçoit les messages transmis. C'est d'autant plus difficile si la conférence se déroule à dis-

tance, avec des participants présents par ordinateurs interposés. Pour y parvenir, il faut permettre à l'intervenant d'avoir un retour de l'impact de ses propos sur son audience.

Nous avons constaté que la problématique de la reconnaissance d'affect, ne peut être résolue convenablement sans recourir aux techniques de reconnaissance d'actions. Ce n'est pas le type d'actions qui nous intéresse ici, mais plutôt ses caractéristiques de mouvement. Ainsi, une étude du processus de reconnaissance est réalisée, couvrant les techniques de représentation, de classification et de fusion d'informations. A travers cet article, dont les contributions se déclinent principalement autour de ces trois axes, nous souhaitons étudier l'apport informationnel issu de l'étude du mouvement de la personne dans sa globalité (mouvement de la tête, mouvement des bras) par rapport à l'augmentation de la précision de la détection des émotions et à la quantification de leurs intensité.

L'article est organisé comme suit. Dans la Section 2, nous présentons brièvement l'état de l'art. La méthodologie proposée organisée en trois niveaux est décrite dans la Section 3. Nous présentons ensuite les résultats obtenus en utilisant la base SEMAINE [MVCP10] dans la Section 4. Afin de valider nos contributions, nous avons choisi un sous-ensemble de la base SEMAINE (utilisée dans les challenges AVEC) car elle présente des personnes assises face à la caméra en conversation avec un agent de manière très similaire à une vidéo-conférence. Les expressions faciales et corporelles sont spontanées car aucune instruction spécifique n'a été transmise aux participants. Enfin, nous concluons par discuter les pistes ouvertes par notre présente contribution dans la Section 5.

2. État de l'art

La définition de l'émotion est difficile à formaliser et dépend du contexte d'usage. Ainsi, là où un neurologue par exemple, sera attaché à des notions de facteurs somatiques ou d'activation cérébrale, un sociologue aura une vision bien plus globale et déterminera des valeurs liées à des paramètres sociaux. Selon les psychologues, certaines émotions de base sont universellement reconnues. Les descripteurs les plus couramment utilisés sont les six émotions de base : la colère, le dégout, la peur, la joie, la surprise, et la tristesse. La plupart des systèmes proposés tentent de reconnaitre un ensemble de prototype d'expressions émotionnelles sur le visage. Pour décrire les changements subtils du visage, le système d'action faciales (FACS) proposé par Ekman [HEF02] est largement utilisé. Une alternative à cette représentation catégorique est l'utilisation de trois dimensions : "agréable ou non agréable" (Valence), "réveil ou soumission" (Arousal) et "tension ou relaxation" (Stance). La Figure 1 illustre l'espace Valence/Arousal labellisé par Russel [Rus80].

L'essentiel de la littérature sur les émotions a été consacrée à l'étude d'une seule modalité qui est le visage. La plupart des travaux existants combinant différentes modalités sont orientés principalement sur les signaux vocaux et l'expression du visage [JSC*04]. L'émotion communiquée à travers les expressions corporelles a souvent été négligée. Le langage du corps est une forme de communication non verbale utilisé pour véhiculer certains messages qu'il est possible d'interpréter par l'observation attentive des gestes, expressions faciales et bien d'autres signaux et mouvements corporels. Van den Stock et al. [VDHdG11] ont constaté que la reconnaissance des expressions faciales est fortement influencée par l'expression corporelle. La plupart des signaux n'étant pas universels, ils doivent être interprétés en fonction du contexte, de l'émetteur, du récepteur, de la culture, etc. [JBS*09].

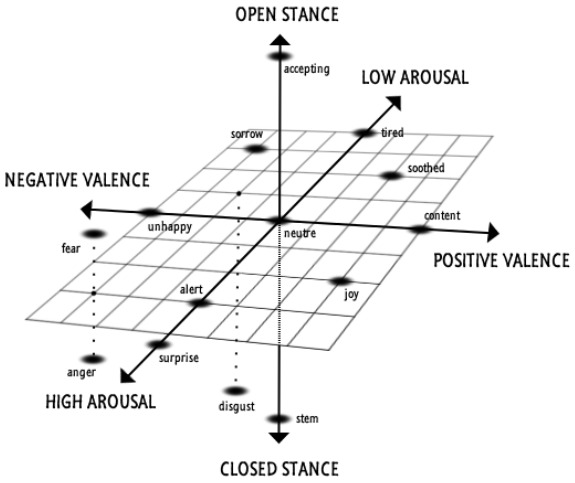


Figure 1: *Espace Valence/Arousal labellisé par Russel [Rus80].*

En analysant le mouvement global du corps (détendu ou contracté) et le mouvement du visage, des mains et des épaules, Gunes et Piccardi [GP05] arrivent à identifier les six états affectifs de base. Plus récemment, Pantic et al. [GNP11] analysent les mouvements des épaules et du visage en complément du canal audio afin d'identifier le niveau de valence et d'arousal pour en déduire l'état affectif.

En complément du geste, la posture exprime l'état psychologique : le degré d'assurance, de concentration, de maîtrise et de la situation en général. Les expériences présentées par Coulson [Cou04] suggèrent que l'interprétation de la posture du corps est comparable à la reconnaissance de la voix, et certaines postures traduisent les même émotions aussi bien que les expressions du visage. Castellano et al. [CVC07] ont présenté une approche pour la reconnaissance de l'émotion basée sur l'analyse des mouvements du corps et de l'expressivité du geste. Ils ont utilisé des données telles que l'amplitude, la vitesse et la fluidité des mouvements pour caractériser 4 émotions de base. La tristesse est représentée par une vitesse et une fluidité des mouvements lente alors que la joie est représentée par des valeurs importantes par exemple. Chen et al. [CTLM13] ont combiné MHI-HOG et Image-HOG à travers une méthode de normalisation temporelle pour décrire la dynamique du visage et des gestes du corps pour estimer l'état affectif. Hadjerci et al. [HLBD14] ont quant à eux utilisés l'information présente dans le mouvement pour estimer l'état affectif dans chacune des quatre dimensions.

On s'intéresse à l'étude des flux vidéo dans un système multi-canal, et principalement à l'étude du visage et du corps. En effet, le langage du corps est une forme de communication non verbale qui permet de véhiculer certains messages qu'il est possible d'interpréter par l'observation attentive de quelques zones bien précises. Il implique des gestes, des expressions faciales et bien d'autres signaux et mouvements corporels. Tous ces éléments font partie intégrante de la communication. De ces données, nous désirons identifier les états affectifs selon plusieurs dimensions, tout particulièrement la valence et l'arousal. Nous choisissons une représentation dimensionnelle, par rapport à une représentation discrète, car par rapport aux domaines ciblés où la durée de l'interaction est relativement longue, il est plus utile de connaître l'évolution de l'état de l'individu, que d'observer des réactions immédiates telles que la surprise, la joie, etc. Pour cela, nous explorons la caractérisation du mouvement à base de modèles de direction et de magnitude.

3. Méthodologie

Nous proposons une méthodologie organisée en trois niveaux pour la reconnaissance d'émotions : (i) Le bas niveau permet d'extraire certaines informations grâce à l'application des techniques de traitement d'images sur les flux vidéo pour en extraire les points caractéristiques, les zones en mouvement, etc. (ii) Le niveau intermédiaire englobe les descripteurs calculés à partir des caractéristiques de bas niveau tels que la trajectoire de déplacement, la vitesse moyenne, la direction moyenne du mouvement, etc. (iii) Le niveau sémantique dépend entièrement du domaine d'application. Son but est de reconstituer à partir des données du niveau intermédiaire des résultats sur l'analyse du comportement humain qui sont compréhensibles par les utilisateurs. La Figure 2 représente sous la forme d'une pyramide notre approche à trois niveaux.

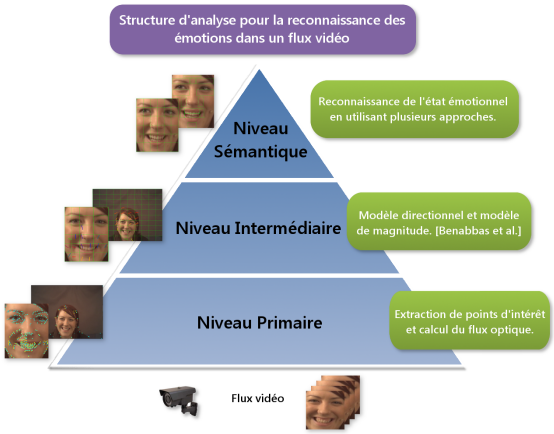


Figure 2: *Schéma en trois niveaux de l'approche de reconnaissance d'état affectif.*

3.1. Extraction des caractéristiques

Cette étape a pour but de quantifier le mouvement à partir des vecteurs de flux optique, afin d'estimer le modèle directionnel et le modèle de magnitude. Nous avons choisi d'utiliser le détecteur de coins de Harris [HS88] pour localiser les points d'intérêt. L'algorithme d'extraction des points d'intérêt de Harris est réputé pour son invariance à la rotation, au changement d'échelle, à la variation de luminosité et aux bruits dans les images. L'algorithme est rapide, ce qui convient aux applications temps réel. Il est aussi déterministe dans le sens où il retourne toujours les mêmes points d'intérêt pour une image donnée en gardant les mêmes paramètres pour l'algorithme. Ensuite, nous appliquons aux points d'intérêt la méthode d'estimation des vecteurs du flux optique KLT [LK81]. Cet algorithme nécessite comme paramètres les pixels de la première image dont nous souhaitons estimer le déplacement. Comme décrit par Baker et Matthews [BM04], l'algorithme recherche pour chaque point présent sur la première image, le point appartenant à la fenêtre de recherche qui lui correspond sur la deuxième image et qui minimise l'équation suivante :

$$\sum_{x,y}[T(x,y) - I(W(x,y;p))]^2 \quad (1)$$

où T est l'apparence du point dont on cherche la correspondance dans la deuxième image, I est la première image, (x,y) un point qui appartient à la fenêtre de recherche de correspondance, W est l'ensemble des transformations envisagées (dans ce cas, la translation) entre la première et la deuxième image et p représente l'ensemble des paramètres de la transformation.

Dans cette étude, nous nous intéressons à deux canal spécifiques : le visage et le corps. Pour chaque canal, nous employons des techniques spécifiques pour choisir les primitives (détection des points de Harris).

Concernant le visage, une étape de détection et de normalisation est nécessaire avant d'identifier les points d'intérêts. Nous appliquons l'algorithme de Viola et Jones [VJ04] pour détecter le visage dans les images. Il est ensuite nécessaire de normaliser les visages afin d'obtenir des modèles cohérents. En effet, le mouvement doit être calculé dans le même repère sur le visage, sinon les régions du visage ne sont plus agencées de la même manière entre deux images. L'algorithme de Danisman et al. [DBID10] s'appuie sur la détection des yeux afin de corriger l'orientation et la normalisation du visage, et permet de suivre le déplacement des yeux dans les images suivantes.

Il faut savoir que le visage n'est pas toujours en mouvement et que l'amplitude de ces mouvements n'est pas forcément constante. Nous cherchons à obtenir des points d'intérêt de qualité, c'est à dire sur des traits marquants du visage et avec un nombre adéquat en fonction de la région analysée.

L'étude du mouvement dans un visage est souvent liée aux contours. En effet, les régions dénuées de contours comme les joues ne sont pas porteuses d'informations. De ce fait, il est intéressant d'extraire les contours du visage pour analyser les mouvements. Pour cela, nous utilisons l'algorithme de Canny [Can86] pour extraire les contours. Parfois, des pré-traitements sont nécessaires pour augmenter la précision des contours (flou gaussien, égalisation d'histogramme, etc.).

Nous proposons une division de l'image en une grille de *M*x*N* blocs pour augmenter le niveau de précision. Ça nous permet également de couvrir autant que possible et de manière homogène le visage et de réduire fortement le temps de calcul de l'appariement des points. La taille de ces blocs et le nombre de points d'intérêt par bloc influent sur la précision du système. La sélection de ces paramètres sera étudiée dans la Section 4. La Figure 3 représente le processus d'extraction des points d'intérêt du visage.

Figure 3: *Processus d'extraction des points d'intérêts et des flux optique sur le visage.*

Afin de déterminer le nombre idéal de points d'intérêt sur le visage, nous calculons le nombre de pixels appartenant au contour dans chaque bloc et nous en déduisons le nombre de points caractérisant le mouvement. Nous appliquons ensuite le calcul du flux optique sur l'image filtrée par Canny et sur l'image suivante en niveau de gris. Cette technique permet de ne perdre aucun mouvement au niveau des contours.

Pour caractériser le mouvement du haut du corps, la méthode diffère quelque peu. En effet, la détection des contours est moins pertinente à cause de l'arrière plan. De ce fait, nous appliquons un algorithme d'extraction de silhouette (estimation des couleurs de l'arrière plan pour les enlever) afin d'obtenir uniquement les mouvements liés au corps. De la même manière, nous divisions l'image en *M*x*N* blocs afin d'augmenter la précision et de calculer le nombre de points d'intérêt par région. Ici, nous déterminons le nombre de points en fonction de la taille d'un bloc. Plus les bloc sont petits, plus le nombre de points d'intérêt par bloc diminue, ce qui a pour effet de réduire le bruit produit par un surnombre d'informations. Ensuite, nous appliquons l'algorithme du calcul des flux optique pour créer nos vecteurs de mouvement. La Figure 4 illustre le processus d'extraction des flux optique du corps.

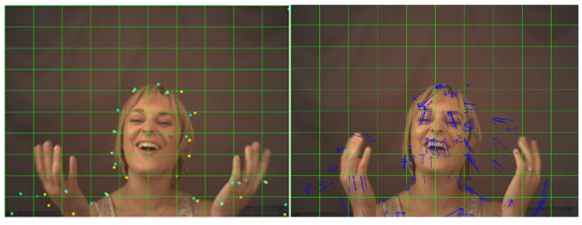


Figure 4: *Processus d'extraction des flux optique du corps.*

3.2. Modèle de direction et de magnitude

Cette étape estime le modèle directionnel et le modèle de magnitude pour l'intégralité de la séquence. Ces modèles serviront à constituer le modèle correspondant à l'évolution de l'état affectif contenu dans la séquence. Pour estimer nos modèles, nous adaptons les travaux de Benabbas et al. [BALD11], destinés aux actions, pour modéliser des changements d'états émotionnels.

La problématique de la reconnaissance d'état affectif, ne peut être résolue convenablement sans recourir aux techniques de reconnaissance d'action. Ce n'est pas le type d'action spécifique qui nous intéresse ici, mais plutôt ses caractéristiques de mouvement. Il est important d'étudier la différence entre l'étude des actions et l'étude des émotions. Pour l'un, nous recherchons à caractériser le mouvement alors que pour l'autre, nous nous intéressons plus particulièrement à l'étude du geste.

A la différence des actions, un geste est un mouvement du corps qui souligne une idée, révèle une pensée ou exprime une émotion. Les mouvements relatifs aux gestes sont plus difficiles à identifier car ils sont moins amples, plus rares et difficilement répétables. Il est donc nécessaire d'adapter les travaux liés à la détection d'actions pour parvenir à caractériser des gestes.

Après avoir calculé les vecteurs de mouvement dans chaque bloc, un algorithme de regroupement des données circulaires est appliqué aux orientations des vecteurs de flux optique. L'ensemble des *M*x*N* distributions circulaires associées est appelé "modèle directionnel". Par analogie, nous regroupons les magnitudes des vecteurs du flux optique dans chaque bloc grâce à des mélanges gaussiens. L'ensemble des mélanges gaussiens estimés représente le modèle de magnitude. Nous appliquons un seuil d'acceptation pour filtrer les données, ce qui permet d'enlever les mouvements, trop petits ou trop grands, qui ne caractérise pas un geste. La Figure 5 représente le modèle de direction et de magnitude construits à partir des flux optiques.

Figure 5: *Création du modèle de direction (b) et du modèle de magnitude (c) à partir des flux optiques (a).*

3.3. Reconnaissance

Cette étape a pour but de reconnaitre les changements d'états affectifs dans une vidéo en comparant les modèles de direction et de magnitude obtenus avec les séquences vidéos de référence. Afin d'analyser les évolutions des états affectifs, nous nous intéressons à l'étude du contexte. En effet, un état émotionnel se caractérise par une suite d'évènements

(gestes, expressions faciales et corporelles). En effectuant des regroupements sur les modèles de direction et de magnitude, nous voulons identifier une suite de mouvements afin de reconnaitre l'état affectif. Pour cela, différentes solutions sont envisageables, notamment la somme des modèles sur un intervalle prédéfini ou bien la sélection des plus grandes amplitudes dans chaque bloc de notre division. Il existe plusieurs techniques pour reconnaitre un état affectif à partir des modèles de référence. Nous en avons identifié trois.

La première solution consiste à comparer le modèle d'une séquence avec les modèles associés aux séquences de référence en utilisant une mesure de distance. L'état affectif associée au modèle ayant la distance la plus petite par rapport au modèle d'une séquence est retenue. Cette solution a l'avantage d'être rapide, et adaptée à une analyse locale (visage) et globale (corps).

La deuxième solution s'inspire des travaux de Gizatdinova et Surakka [GS07], où ils divisent le visage en 13 ROIs (Region of Interest) pour détecter les émotions. Une étude comparative [PRW*11] montre que cette solution permet d'améliorer le taux de reconnaissance par rapport à une division en *M*x*N* blocs. Cependant, cette technique s'applique uniquement au visage car à l'heure actuelle il n'existe pas de système équivalent aux AUs pour le corps.

Suite à une étude approfondie sur les unités d'action et l'évolution de l'état affectif, nous envisageons d'étendre les travaux de Gizatdinova et Surakka, en adaptant la division du visage en fonction des AUs identifiées comme pertinent. Nous voulons analyser des suites d'AUs et anticiper l'évolution de l'état affectif en fonction de l'ordre d'apparition des mouvements au sein du visage.

La troisième solution regroupe les deux solutions précédentes. L'idée est d'appliquer la solution globale sur le corps et la solution locale sur le visage. Grâce à cela, nous augmentons le taux de reconnaissance sur le visage tout en gardant les informations relatives au mouvement global.

Étant donné, que la solution deux et trois dépendent fortement de la normalisation du visage, qui fait l'objet de nos récents travaux, nous nous décidons d'étudier la première solution aussi bien sur le visage, que sur le haut du corps. Cette solution nous permet d'obtenir des premiers résultats pour analyser la pertinence de l'étude du mouvement dans la reconnaissance d'état affectif.

4. Résultats expérimentaux

Nous présentons dans cette section les résultat de l'expérimentation de notre approche sur un sous-ensemble de la base de données SEMAINE. Cette base permet d'étudier les signaux sociaux naturels qui se produisent dans des conversations entre humains et agents artificiels. Les vidéos ont été enregistrées à une fréquence de 50 images par secondes avec une résolution de 780x580 pixels. La base est composée de 31 vidéos d'apprentissage et de 32 vidéos utilisés pour les tests. La particularité de cette base réside dans le fait qu'elle est annotée en continue (traces) par au moins deux évaluateurs. L'analyse est effectuée sur l'évolution des dimensions affectives telles que l'activité (Arousal), l'anticipation, la puissance et la valence. Notre sous-ensemble d'étude est composé de 11 vidéos d'entrainement et de 8 vidéos de test provenant de deux sujets différents.

Nous nous intéressons tout particulièrement aux variations des émotions, c'est à dire aux changements de valeurs de la valence ou de l'arousal. Afin de catégoriser une émotion, nous nous inspirons du modèle de Russel, ce qui nous donne les quatre classes suivantes : Arousal +/ Valence +, Arousal +/ Valence -, Arousal -/ Valence +, Arousal - / Valence -. Chaque classe correspond à une augmentation ou une diminution de la valeur initiale, c'est à dire qu'un modèle se situant dans la classe Arousal +/Valence + n'a pas forcement une valence et une activité positives mais ses valeurs ont augmenté par rapport au modèle précédent.

Afin de valider nos résultats, nous optons pour la classification avec les SVM [CV95] puis nous utilisons LIBSVM [CL11] pour optimiser les paramètres du Kernel par rapport aux données dont nous disposons. Chaque modèle de direction et de magnitude est représenté par un vecteur. Pour cela, nous déterminons la classe du modèle en fonction de l'évolution des valeurs de l'arousal et de la valence, ce qui nous donne un label variant de 1 à 4. Concernant les index, ils correspondent aux blocs de l'image et les values, à leurs valeurs respectives. Nous décidons de réunir les données de direction et les données de magnitude dans le même classifieur afin d'obtenir une meilleure précision lors de la classification. De ce fait, tous les index impairs correspondent à la valeur de la magnitude et les index pairs à la direction. La construction de ce vecteur est représentée sur la Figure 6.

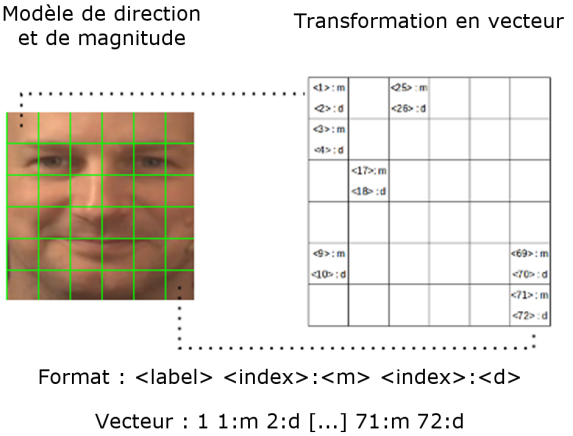


Figure 6: *Construction de vecteur à partir d'un modèle de direction et de magnitude*

La taille des vecteurs varie en fonction du nombre de blocs dans le modèle. De plus, il est possible d'ajouter d'autres données comme par exemple la puissance. Pour cela, il faut associer 3 index par bloc. Quant au nombre de vecteurs, il dépend fortement du nombre de divisions et du seuil d'acceptation appliqué aux magnitudes. En effet, plus le seuil est bas et le nombre de division élevé, plus l'algorithme construit de modèles. La Figure 7 présente le taux de reconnaissance avec une approche globale sur des visages non normalisés.

Sur la Figure 7, nous remarquons que le taux de reconnaissance varie en fonction de plusieurs paramètres. La précision du modèle est influencée par le nombre de blocs. En effet, plus la division est grande, plus il y a de points caractéristiques. Cependant, un nombre trop important de blocs peut réduire la précision du modèle car les données ne sont plus pertinentes pour calculer le mouvement. La taille des blocs dépend de la dimension des images. Lorsque nous appliquons la même division sur des images de tailles différentes, nous constatons une augmentation du taux de reconnaissance.

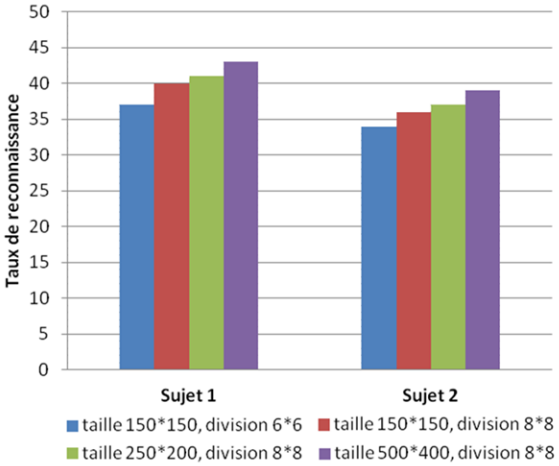


Figure 7: *Taux de reconnaissance sur le visage.*

En contrepartie, plus la taille de l'image est grande, plus le temps de calcul est long. De ce fait, il faut trouver le bon compromis entre précision et temps d'exécution par rapport à un contexte pratique spécifique. Etant donné que la normalisation n'est pas encore appliquée sur les visages, cela peut expliquer pourquoi le taux de reconnaissance n'est pas très élevé. La différence entre le Sujet 1 et le Sujet 2 est probablement liée au nombre de vidéos d'entrainement qui est plus conséquent chez le Sujet 1.

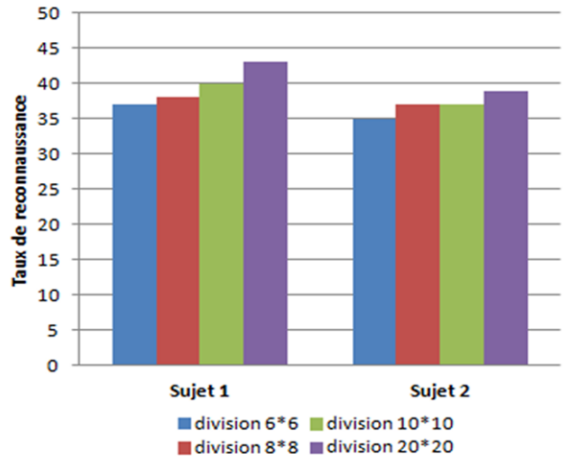


Figure 8: *Taux de reconnaissance sur le haut du corps.*

Nous faisons la même analyse sur les mouvements du corps et nous obtenons les résultats présentés sur la Figure 8. Nous remarquons que le taux de reconnaissance des mouvements du corps est plus ou moins équivalent à celui lié au visage. Au cours de ces conversations, les sujets ne bougent pas beaucoup ce qui implique que le classifieur n'arrive pas à dissocier les classes. De plus, les résultats du Sujet 2 sont inférieurs à ceux du Sujet 1 car ses vêtements se confondent avec l'arrière plan, de ce fait, nous obtenons uniquement les mouvements liés à sa tête.

4.1. Discussions

Il est important de préciser que la reconnaissance affective multicanaux ne vise pas à remplacer les expressions faciales ou les expressions corporelles, au lieu de cela, le but est d'explorer divers canaux communicatifs plus profonds et plus complets afin d'obtenir des corrélations relatives à l'état affectif.

Le langage du corps implique gestes, expressions faciales et bien d'autres signaux et mouvements corporels. Tous ces éléments font partie intégrante de la communication. Or, la complexité de chaque personnalité fait qu'il existe une multiplicité de gestes qui trahissent chacun d'entre nous. Cependant, le sens des gestes peut être interprété de plusieurs façons.

L'étude du mouvement permet donc d'identifier et de reconnaitre ces gestes. Cependant, l'interprétation des résultats n'est pas évidente. Ce problème réside dans le fait, qu'il n'existe pas une définition formelle du terme "émotion". Il est d'autant plus difficile de construire une base d'apprentissage lorsqu'on n'arrive pas à représenter l'objet de l'étude. Cette étude montre qu'il existe un nombre important de paramètres qui influencent le taux de reconnaissance. En effet, la dimension, la résolution de l'image et la division, sont des paramètres que l'on ne peut définir en amont car ils varient en fonction de la vidéo. De plus, la précision des modèles repose essentiellement sur la robustesse des algorithmes de normalisations et des descripteurs.

Au fil des résultats, nous obtenons un taux de reconnaissance meilleur sur le visage par rapport au corps. D'après ces données, nous considérons que le mouvement du corps est un complément d'information à celui du visage. Il permet de donner une intensité à l'état affectif, comme certaines AUs sur le visage.

Enfin, il existe plusieurs approches permettant d'identifier un état affectif. Il est possible de calculer un modèle à un instant précis de la vidéo, ou bien sur un intervalle. Durant l'étude nous avons constaté qu'il y a parfois des variations dans les courbes alors qu'aucun changement n'apparait dans le flux vidéo. En effet, l'état émotionnel dépend également du contexte, ce qui nous amène à construire des modèles de direction et de magnitude sur des intervalles et à identifier les mouvements déclencheurs en analysant les suites de mouvement. Bien entendu, les solutions sur les intervalles sont très variées, ce qui prouve une fois de plus la complexité et la richesse de cette étude.

5. Conclusion

Dans cet article, nous avons proposé une approche organisée en trois niveaux pour la reconnaissance d'états affectifs multi-canal. Notre système synthétise les mouvements du visage et du corps par des modèles de direction et de magnitude construits à partir des flux optiques. Les résultats expérimentaux permettent de vérifier l'apport informationnel issu du mouvement dans la reconnaissance d'états affectifs et montrent que cette solution est adaptable dans une approche globale ou locale. Dans nos futurs travaux, nous allons améliorer la normalisation des visages ce qui nous permettra d'analyser l'ensemble de la base SEMAINE. Nous analyserons également l'apport de la fusion multi-canal et de l'analyse du contexte dans l'amélioration du taux de reconnaissance.

Références

[AR92] AMBADY N., ROSENTHAL R. : Thin slices of expressive behavior as predictors of interpersonal consequences : A meta-analysis. *Psychological Bulletin. Vol. 111*, Num. 2 (1992), 256–274.

[BALD11] BENABBAS Y., AMIR S., LABLACK A., DJERABA C. : Human action recognition using direction and magnitude models of motion. In *International Conference on Computer Vision Theory and Applications (VISAPP)* (2011).

[BM04] BAKER S., MATTHEWS I. : Lucas-Kanade 20 years on : A unifying framework. *International Journal of Computer Vision (IJCV)* (2004).

[Can86] CANNY J. : A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (1986).

[CL11] CHANG C.-C., LIN C.-J. : LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology. Vol. 2* (2011).

[Cou04] COULSON M. : Attributing emotion to static body postures : Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior* (2004).

[CTLM13] CHEN S., TIAN Y., LIU Q., METAXAS D. : Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing (IVC). Vol. 31*, Num. 2 (2013), 175–185.

[CV95] CORTES C., VAPNIK V. : Support-vector networks. *Machine Learning. Vol. 20* (1995), 273–297.

[CVC07] CASTELLANO G., VILLALBA S. D., CAMURRI A. : Recognising human emotions from body movement and gesture dynamics. *Affective Computing and Intelligent Interaction* (2007).

[DBID10] DANISMAN T., BILASCO I. M., IHADDADENE N., DJERABA C. : Automatic facial feature detection for facial expression recognition. In *5th International Conference on Computer Vision Theory and Applications (VISAPP)* (2010).

[GNP11] GUNES H., NICOLAOU M., PANTIC M. : Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing (TAC)* (2011).

[GP05] GUNES H., PICCARDI M. : Affect recognition from face and body : Early fusion vs. late fusion. *Systems, Man and Cybernetics* (2005).

[GS07] GIZATDINOVA Y., SURAKKA V. : Automatic detection of facial landmarks from au-coded expressive facial images. *International Conference on Image Analysis and Processing (ICIAP)* (2007).

[HEF02] HAGER J., EKMAN P., FRIESEN W. : The facial action coding system : A technique for the measurement of facial movement. *San Francisco : Consulting Psychologist* (2002).

[HLBD14] HADJERCI O., LABLACK A., BILASCO I. M., DJERABA C. : Affect recognition using magnitude models of motion. In *20th International Conference on MultiMedia Modeling (MMM)* (2014).

[HS88] HARRIS C., STEPHENS M. : A combined corner and edge detector. *Alvey Vision Conference* (1988).

[JBS*09] JACK R., BLAIS C., SCHEEPERS C., SCHYNS P., CALDARA R. : Cultural confusions show that facial expressions are not universal. *Current Biology* (2009).

[JSC*04] JING T. M., SCHMIDT X. K., COHN J., REED L., AMBADAR Z. : Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (2004).

[KCY00] KANADE T., COHN J., YINGLI T. : Comprehensive database for facial expression analysis. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (2000).

[LK81] LUCAS B., KANADE T. : An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence (IJCAI)* (1981).

[MVCP10] MCKEOWN G., VALSTAR M., COWIE R., PANTIC M. : The SEMAINE corpus of emotionally coloured character interactions. In *IEEE International Conference on Multimedia and Expo (ICME* (2010).

[PRW*11] POPA M., ROTHKRANTZ L., WIGGERS P., BRASPENNING R., SHAN C. : Facial action units recognition - a comparative study. *IEEE Transactions on Multimedia special issue on Multimodal Affective Iinteraction* (2011).

[Rus80] RUSSELL J. A. : A circumplex model of affect. In *Journal of Personality and Social Psychology, Vol 39(6)* (1980).

[VDHdG11] VAN DEN STOCK J., DE JONG S., HODIAMONT P., DE GELDER B. : Perceiving emotions from bodily expressions and multisensory integration of emotion cues in schizophrenia. *Social Neuroscience. Vol. 6*, Num. 5-6 (2011), 537–547.

[VJ04] VIOLA P., JONES M. J. : Robust real-time face detection. *International Journal of Computer Vision (IJCV). Vol. 57*, Num. 2 (2004), 137–154.

Caractérisation locale des changements de texture pour la reconnaissance d'expressions faciales spontanées

Walid Adaidi, Adel Lablack, Ioan Marius Bilasco

Laboratoire d'Informatique Fondamentale de Lille, Université de Lille 1, France

Résumé

Malgré les avancées récentes, la reconnaissance des émotions et des expressions faciales reste un challenge intéressant. Dans cet article, une approche permettant la reconnaissance d'expressions faciales spontanées grâce à une représentation appropriée des traits du visage sur des flux vidéo et des images statiques est proposée. Une mesure sensible aux changements dans les traits du visage est utilisée dans des régions d'intérêt identifiées pour détecter la présence de chaque expression de base. L'expérimentation a été réalisée sur un ensemble de données standard composées de vidéos et d'images statiques et a montré des résultats prometteurs.

Recognizing human facial expression and emotion by computer is an interesting and challenging problem. In this paper, a method for recognizing spontaneous facial expressions through an appropriate representation of facial features from relevant face regions displayed in video streams and still images is proposed. A measure that is sensitive to facial movements is used in predefined regions of interest to detect the basic emotions. The experimentation has been performed on a standard dataset composed of video streams and static images and has showed promising results.

Mots clé : Reconnaissance d'expressions faciales spontanées, approche locale, régions d'intérêt

1. Introduction

La reconnaissance automatique des expressions faciales est un sujet de recherche actif. Les techniques proposées actuellement pour la détection et le suivi du visage, l'extraction des caractéristiques du visage et les méthodes utilisées pour la classification des expressions sont plus robustes qu'auparavant. L'expression faciale peut être utilisée comme un module important dans les interactions homme-machine ou pour étudier le comportement des personnes. L'interprétation des expressions faciales diffère d'un domaine d'application à un autre. C'est une tâche difficile qui permet au système d'être réactif et d'améliorer l'expérience utilisateur.

Dans un magasin par exemple, un retour positif (un sourire) peut être interprété comme un signe d'intérêt, alors qu'une grimace pourrait être interprétée comme un signe de répulsion. Selon la réactivité souhaitée, le système pourrait présenter plus de détails sur un produit ou changer le produit affiché à l'utilisateur. Dans les applications de e-learning, il faut considérer une échelle temporelle plus large car l'interaction de l'utilisateur avec le système ou bien avec l'enseignant est censée se dérouler en continue. Dans ce cas, des états tels que l'intérêt de l'utilisateur, l'incompréhension, ou la frustration peuvent être considérer comme des actions. Ces deux exemples montrent différentes façon de représenter l'état émotionnel de l'utilisateur.

Dans la littérature deux approches principales pour représenter les expressions sont utilisés : une représentation discrète en catégories introduite par Ekman [EF78] qui utilise six expressions faciales universelles que sont la colère, le dégoût, la peur, la joie, la tristesse et la surprise. La représentation dimensionnelle est une alternative et permet de caractériser un état affectif en termes de dimensions latentes [CD10] telles que l'évaluation, l'activation, le contrôle, la puissance, etc.

En présence d'expressions spontanées et non actées dans les domaines applicatifs visés, nous nous sommes intéressés dans cet article à la reconnaissance d'expressions faciales spontanées dans des environnements non contrôlés. En effet, l'identification des unités d'actions (AUs) à l'aide d'un modèle et leurs suivis est difficile sur le visage à cause du bruit. Ce bruit est souvent confondu avec des mouvements de très faible amplitude. Par ailleurs, dans une situation d'interactions spontanées les règles FACS proposées par Ekman [EF78] ne couvrent pas l'intégralité des situations.

A l'image de Mavadati et al. [MMB*13] qui s'intéressent à la caractérisation des expressions faciales spontanées, nous commençons par étudier le lien entre les changements de texture dans le visage et les expressions de base. En effet, nous adoptons une analyse des changements intervenus sur

des régions spécifiques de visage afin de s'affranchir du délicat problème d'extraction directe des AUs dans un environnement non contrôlé, nous adoptons une analyse des changements intervenus sur des régions spécifiques de visage. Une étude préalable entre les changements observés et les expressions a été conduite sur la base DISFA [MMB*13]. Cette base est constituée de vidéos de 27 personnes avec une annotation manuelle de l'intensité de 12 AUs et de la position de 66 points de contrôle sur le visage.

L'article est organisé comme suit. Dans la section 2, nous présentons brièvement l'état de l'art. La méthodologie proposée organisée en deux étapes est décrite dans la section 3. Nous présentons ensuite les résultats obtenus en utilisant les bases DISFA et KDEF [LFO98] dans la section 4. Enfin, nous concluons par discuter les pistes ouvertes par notre présente contribution dans la section 5.

2. État de l'art

Le domaine de la reconnaissance d'émotions a été étudié activement au cours de ces dernières années. En général, le but des systèmes proposés est de reconnaître des classes d'expressions ou bien l'état d'une dimension affective. Dans la littérature, les techniques utilisées pour détecter des expressions dans des flux vidéo peuvent être classés en deux ensembles : (i) les approches géométrique pour détecter et suivre des points caractéristiques du visage qui sont ensuite utilisés en entrée dans la classification, (ii) les approches d'apparence qui utilisent le mouvement et le changement de texture [SRS*11].

Les approches géométriques s'appuient essentiellement sur le système Facial Action Coding System (FACS) qui permet de mesurer les changements subtils dans l'apparence du visage causées par des contractions des muscles faciaux en associant une action unitaire à chaque mouvement musculaire d'une partie du visage [EF78]. Gonzalez et al. [GSEV11] ont appliqué adaboost à un ensemble de caractéristiques extrait dans les régions du visage permettant de reconnaitre les AUs. Popa et al. [PRW*11] ont mené une étude sur la reconnaissance des AUs. Ils se sont appuyés sur le principe que les yeux, le nez et la bouche contiennent beaucoup d'informations qu'il faut ensuite affiner en se basant sur un AAM [CET01]. Ils utilisent le flux optique pour extraire les informations contenues dans les régions du visage mais cette approche est sensible aux mouvements. Lablack et al. [LDBD14] se sont intéressés uniquement aux émotions négatives et ont proposé une approche locale autour de la région englobant l'AU4.

Des méthodes globales telles que celle présentée par Danisman et al. [DBMD13] qui utilisent un perceptron multicouche pour reconnaitre la joie semblent plus robustes aux bruits mais ne sont pas adaptées à toutes les expressions.

Nous proposons une approche locale pour détecter les expressions faciales tout en gardant les bénéfices des méthodes globales. Ainsi, nous étudions les changements de texture consécutifs à l'apparition de certaines AUs sur le visage. Nous proposons une analyse spécifique des régions à proximité ou englobant ces AUs sur le visage afin de détecter les changements qui apparaissent en présence d'une expression.

3. Méthodologie

Nous proposons une méthodologie qui permet de caractériser les expressions faciales spontanées de base en s'appuyant sur une analyse des parties locales du visage. Notre proposition se divise en deux étapes importantes : (i) l'étude préalable des régions qui permet d'identifier les régions d'intérêt du visage sur lesquels un changement apparait lors de l'expression d'une émotion par l'utilisateur (ii) extraire les informations caractéristiques des régions obtenues en associant une métrique à chaque expression faciale. La Figure 1 illustre le schéma proposé en deux étapes.

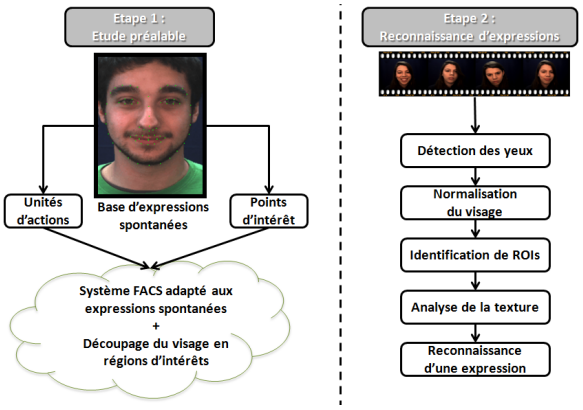
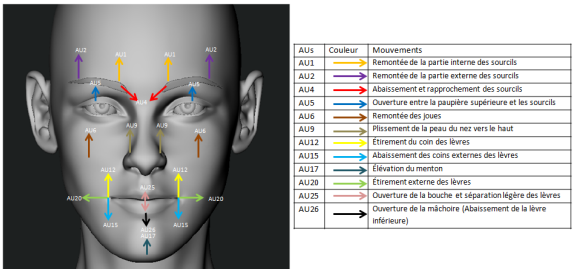


Figure 1: *Le système proposé pour la détection des expressions faciales de base.*

3.1. Étude préalable des régions

Cette étude va nous permettre d'associer à chaque expression de base une région d'intérêt à analyser. Pour cela, nous disposons des coordonnées des points d'intérêts dans la base DISFA [MMB*13]. Les points d'intérêts sont placés tout autour des composants du visage tels que le nez, les sourcils, la bouche et les yeux. Une activation d'une unité d'action (AU) entraine une déformation faciale, donc un déplacement caractérisé par la différence entre les points d'intérêt à l'instant t et à l'instant $t+1$. Nous allons utiliser la combinaison des AUs formant une expression pour définir nos régions d'intérêts. La Figure 2 illustre la position et les mouvements faciaux qui caractérisent les 12 AUs que nous utilisons pour l'identification des régions d'intérêts. Par exemple pour la partie autour des sourcils : l'AU1 correspond à la remontée de la partie interne des sourcils, l'AU2 à la remontée de la partie externe des sourcils, et l'AU4 à l'abaissement et rapprochement des sourcils.

Le système établi par Ekman reste théorique et difficile à appliquer en présence d'expressions faciales spontanées. Nous avons analysés l'activation des AUs sur la base DISFA afin d'identifier et valider le changement des traits du visage en présence d'expressions faciales spontanées. Selon Ekman qui a construit son modèle sur une base d'expressions actées, chaque expression nécessite l'activation de plusieurs AUs. Par exemple, selon Ekman, la surprise est constituée des AU1, AU2, AU5 et AU26 et l'absence d'une seule AU annule le processus de détection. Notre modèle quant à lui se



AUs	Couleur	Mouvements
AU1		Remontée de la partie interne des sourcils
AU2		Remontée de la partie externe des sourcils
AU4		Abaissement et rapprochement des sourcils
AU5		Ouverture entre la paupière supérieure et les sourcils
AU6		Remontée des joues
AU9		Plissement de la peau du nez vers le haut
AU12		Étirement du coin des lèvres
AU15		Abaissement des coins externes des lèvres
AU17		Élévation du menton
AU20		Étirement externe des lèvres
AU25		Ouverture de la bouche et séparation légère des lèvres
AU26		Ouverture de la mâchoire (Abaissement de la lèvre inférieure)

Figure 2: *Localisation des 12 AUs qui caractérisent les expressions de base.*

base sur la probabilité de présence d'une AU ou d'un groupe d'AUs. La surprise peut être détectée avec une combinaison AU1, AU2 et AU5 ou AU1, AU2 et AU26.

Nous avons menée une étude sur la cooccurrence des AUs en présence/absence d'expressions dans la base DISFA. Nous avons constaté par exemple que l'AU26 est souvent présente avec la plupart des autres AUs et se révèle être non pertinente dans la détection des expressions. Le dégout par exemple peut être caractérisé par la présence de l'AU9 avec une grande intensité alors qu'Ekman conditionne sa détection à la présence des AU15 et AU16. Ces observations nous semblent importantes dans l'étude des solutions adaptées pour la détection d'expressions spontanées car au contraire des expressions actées, l'activation de certains AUs requises par le FACS n'est que partielle ou absente. Suite à cette analyse, nous proposons une simplification du modèle classique FACS en réduisant le nombre d'AUs nécessaires à la caractérisation d'une expression faciale. La Figure 3 illustre une représentation en portes logiques des combinaisons des AUs formant une expression faciale spontanée que nous avons retenu en interprétant les données de la base DISFA.

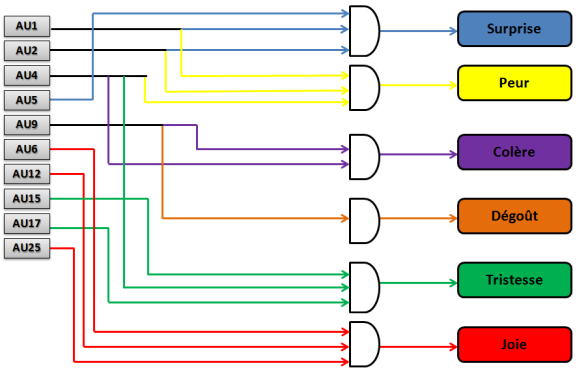


Figure 3: *La représentation en portes logiques des combinaisons des AUs formant un expression faciale.*

Une fois cette association entre les AUs et les expressions établie, nous analysons la position des points d'intérêts en présence/absence de chaque expression. Les zones sont choisies de façon à englober les points d'intérêts lors du déclenchement de l'émotion. Toutefois, nous avons effectuer une normalisation du visage qui permet de compenser les bruits dus à la variation de la position spatiale du visage. En effet, une variation entraîne le déplacement des points d'intérêt et induit donc un bruit dans l'analyse. Nous avons affiner les zones pour obtenir la région minimale qui contiendra tous les points d'intérêts. La Figure 4 illustre un découpage des régions pour les expressions surprise, joie, et dégout.

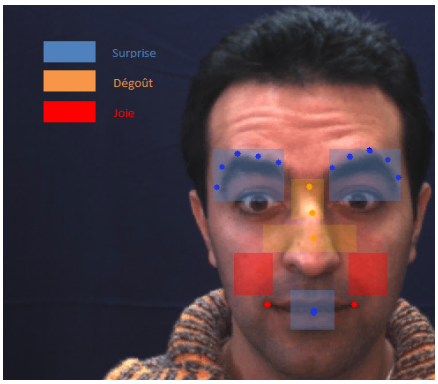


Figure 4: *Découpage du visage en régions d'intérêt.*

3.2. Reconnaissance d'expressions

Après avoir trouvé et affiné les régions correspondant à chaque expression, nous allons analyser ces régions afin d'identifier les changements de texture en présence d'une expression particulière. Cette analyse de la texture est faite en appliquant des filtres tels que le filtre de Gabor ou le filtre LBP. Le filtre LBP possède certaines propriétés qui lui permettent d'être efficace en pratique. Il est robuste aux conditions d'éclairage [AHP04], performant à la sélection de paramètres [OPM02], ne nécessite pas d'initialisation et fonctionne de manière fiable sur des résolutions d'images basses [SGM05].

Nous avons choisi d'appliquer les filtres de Gabor et de LBP avec différents paramètres sur les régions associées à chaque expression. Nous avons ensuite maximisé la valeur de chaque mesure pour obtenir des pics lors de l'activation d'une expression tout en minimisant sa valeur en absence de cette expression.

4. Résultats expérimentaux

Nous avons validé notre approche sur deux types de bases. Une base contenant des vidéos de personnes regardant une webcam et exprimant spontanément différentes expressions. La seconde est composée d'images statiques de personnes exprimant des expressions actées qui nous permet de valider notre étude préliminaire sur la base DISFA.

4.1. La base vidéo DISFA

La Figure 5 présente le résultat obtenu pour la détection de la surprise sur une vidéo de la base DISFA. Le participant a exprimé différentes émotions et les instants où il a exprimé la surprise sont illustrés dans la partie haute de la figure. La partie basse présente les résultats de notre métrique qui n'est pas influencée par les autres expressions puisque les pics surviennent aux moments où la surprise a été annotée sur la vidéo.

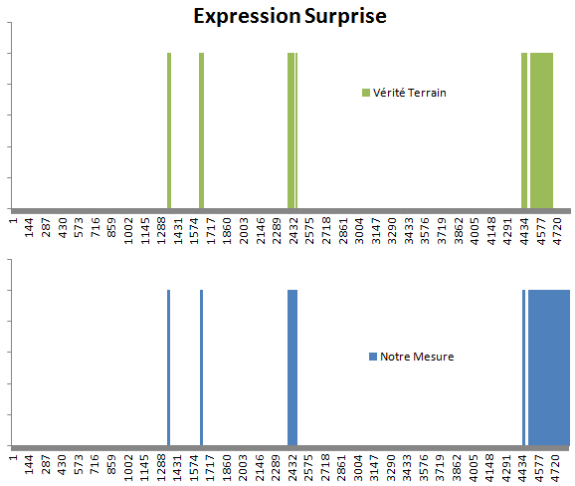


Figure 5: *Détection de la surprise dans une vidéo.*

4.2. La base d'images KDEF

La base Karolinska Directed Emotional Faces (KDEF) [LFO98] contient un ensemble d'expressions faciales actées qui permettent de renforcer la validation de notre étude préalable. La base contient 70 personnes, chacune affichant sept expressions différentes (neutre, joie, colère, peur, dégoût, tristesse, surprise) avec chaque expression photographiée deux fois. Les participants étaient assis à une distance d'environ trois mètres de la caméra. La résolution des images est de 562x762. La Figure 6 présente les résultats obtenus pour les expressions "surprise", "joie" et "dégout" sur la base KDEF. La mesure associée à une expression est considérée détectée correctement lorsque sa valeur est la plus élevée.

Expression	Peur	Colère	Dégoût	Joie	Neutre	Tristesse	Surprise
Joie	1,43%	1,43%	0,00%	**92,14%**	0,71%	4,29%	0,00%
Surprise	7,14%	2,41%	1,43%	0,71%	4,29%	6,43%	**77,86%**
Dégoût	2,14%	21,43%	**70,00%**	0,71%	2,14%	1,43%	2,14%

Figure 6: *Les résultats de l'application de la mesure sur les expressions surprise, joie et dégout de la base KDEF.*

5. Conclusion

Dans cet article, nous avons proposé une approche locale pour caractériser les changements de texture pour la reconnaissance d'expressions faciales spontanées. Nous associons à chaque expression une mesure qui est sensible aux changements dans les traits du visage dans des régions d'intérêt localisées. Cette approche ne nécessite pas de procédés spécifiques d'apprentissage préalable ou une reconnaissance d'un ensemble d'unités d'action. Les résultats sont prometteurs en termes de robustesse aussi bien pour la détection d'expressions faciales actées que spontanées. Dans nos futurs travaux, nous allons essayer d'augmenter le taux de reconnaissance en appliquant le filtre LBP sur des images filtrées par Gabor. On veut également modéliser par ce même principe des régions, la présence/absence des mouvements faciaux discriminant les expressions faciales ayant des unités d'actions similaires.

Références

[AHP04] AHONEN T., HADID A., , PIETIKAINENG M. : Face recognition with local binary patterns. In *European Conference on Computer Vision (ECCV)* (2004).

[CD10] CALVO R., D'MELLO S. : Affect detection : An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing (TAC). Vol. 1*, Num. 1 (2010), 18–37.

[CET01] COOTES T., EDWARDS G., TAYLOR C. : Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). Vol. 23*, Num. 6 (2001), 681–685.

[DBMD13] DANISMAN T., BILASCO I. M., MARTINET J., DJERABA C. : Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron. *Signal Processing. Vol. 93*, Num. 6 (2013), 1547–1556.

[EF78] EKMAN P., FRIESEN W. : *Facial Action Coding System : A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, 1978.

[GSEV11] GONZALEZ I., SAHLI H., ENESCU V., VERHELST W. : Context-independent facial action unit recognition using shape and gabor phase information. In *4th International Conference on Affective Computing and Intelligent Interaction (ACII)* (2011).

[LDBD14] LABLACK A., DANISMAN T., BILASCO I. M., DJERABA C. : A local approach for negative emotion detection. In *22nd International Conference on Pattern Recognition (ICPR)* (2014).

[LFO98] LUNDQVIST D., FLYKT A., OHMAN A. : *The Karolinska Directed Emotional Faces (KDEF)*. Karolinska Institutet, 1998.

[MMB*13] MAVADATI S. M., MAHOOR M. H., BARTLETT K., TRINH P., COHN J. F. : DISFA : A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing. Vol. 4*, Num. 2 (2013), 151–160.

[OPM02] OJALA T., PIETIKÄINEN M., MÄENPÄÄ T. : Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2002), 971–987.

[PRW*11] POPA M., ROTHKRANTZ L., WIGGERS P., BRASPENNING R., SHAN C. : Facial Action Units Recognition - A comparative study. *IEEE Transactions on Multimedia special issue on Multimodal Affective Iinteraction* (2011).

[SGM05] SHAN C., GONG S., MCOWAN P. : Robust facial expression recognition using local binary patterns. In *International Conference on Image Processing (ICIP)* (2005).

[SRS*11] SENECHAL T., RAPP V., SALAM H., SÉGUIER R., BAILLY K., PREVOST L. : Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units. In *International Conference on Automatic Face and Gesture Recognition (FG)* (2011).

Design, Implementation and Simulation of a Cloud Computing System for Enhancing Real-time Video Services by using VANET and Onboard Navigation Systems

K. Hammoudi[1,2] N. Ajam[1,2] M. Kasraoui[1] F. Dornaika[3,4] K. Radhakrishnan[2,*] K. Bandi[2,*] Q. Cai[2,*] S. Liu[2,*]

[1]Research Institute on Embedded Electronic Systems (IRSEEM), IIS Group, Technopôle du Madrillet, St-Etienne-du-Rouvray, France
[2]ESIGELEC School of Engineering, Department of ICT (*MS Students), Technopôle du Madrillet, St-Etienne-du-Rouvray, France
[3] Department of Computer Science and Artificial Intelligence, University of the Basque Country, San Sebastián, Spain
[4] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Résumé

Dans cet article, nous proposons une architecture pour le développement de systèmes de cloud computing nouveaux et expérimentaux. Le système proposé vise à renforcer les capacités de calcul, de communication et d'analyse de services de navigation routière par la fusion de plusieurs technologies indépendantes, à savoir les systèmes de navigation embarqués basés sur la vision, les systèmes de cloud computing de premier plan et les réseaux Ad-Hoc de véhicules (VANET). Ce travail présente nos premières investigations en décrivant la conception d'un système générique global. Le système conçu a été expérimenté à travers deux scénarios de services routiers basés sur la vidéo. En outre, l'architecture associée a été mise en œuvre sur un simulateur à échelle réduite d'un système embarqué véhiculaire. L'architecture développée a été testée dans le cas d'une application routière simulée visant à aider certains services de police. Le but de cette application est de reconnaître et de pister des véhicules et des individus recherchés en temps réel moyennant un système de surveillance formé par des véhicules en circulation. Le travail présenté démontre le potentiel de notre système pour améliorer efficacement et pour diversifier les applications nécessitant des traitements de vidéos en temps réel dans des environnements routiers.

Abstract

In this paper, we propose a design for novel and experimental cloud computing systems. The proposed system aims at enhancing computational, communicational and annalistic capabilities of road navigation services by merging several independent technologies, namely vision-based embedded navigation systems, prominent Cloud Computing Systems (CCSs) and Vehicular Ad-hoc NETwork (VANET). This work presents our initial investigations by describing the design of a global generic system. The designed system has been experimented with various scenarios of video-based road services. Moreover, the associated architecture has been implemented on a small-scale simulator of an in-vehicle embedded system. The implemented architecture has been experimented in the case of a simulated road service to aid the police agency. The goal of this service is to recognize and track searched individuals and vehicles in a real-time monitoring system remotely connected to moving cars. The presented work demonstrates the potential of our system for efficiently enhancing and diversifying real-time video services in road environments.

Keywords: Vehicular Network (VANET), Vehicular Cloud Computing (VCC), Image-based Recognition, Fusion of Multi-source Imagery, Real-time Video Services, Cooperative Monitoring System, Sensor Networks.

1. Introduction and Motivation

In this work, we propose to exploit cloud computing systems for developing real-time road video services from embedded navigation systems and VANETs (Vehicular Ad-hoc NETworks). The proposed systems will have a final objective to be experimented on a vehicle fleet. More particularly, this paper presents the design, the implementation and the simulation parts of a cloud-based recognition system for extending real-time road video services. Indeed, the proposed global generic system will exploit a cloud-based embedded recognition systems and VANET technologies ; on the one hand, for analyzing the road traffic (e.g. ; vehicular or navigation information) and on the other hand, for mutualizing the computational resources as well as for sharing relevant information visually extracted. Notably, the designed system will be useful for identifying dynamical Points Of Interest from embedded cameras (e.g., traffic-based POI) and then sharing the identified POIs to external stakeholders potentially interested (e.g.,

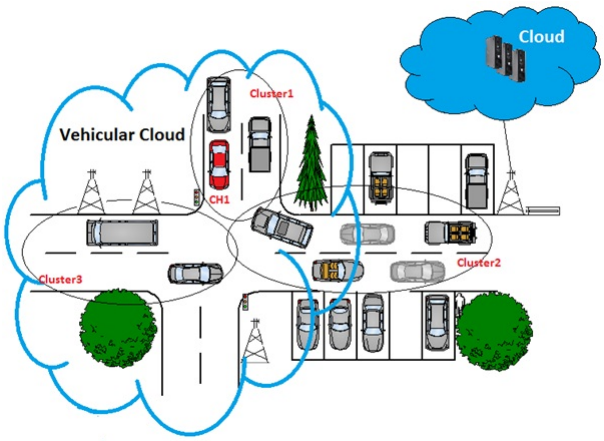


Figure 1: Illustration of the designed cloud computing system.

surrounding vehicles or road agencies).

For instance, these technologies can be exploited for improving the road traffic, the emergency mapping or the citizen security by cooperatively analyzing acquired georeferenced road images. Respectively, we present below some scenarios that will be based on the detection of dynamical POIs :

- **Sc.1 :** a vehicle can detect an available parking area and to transmit its GPS location in a pre-defined neighborhood for informing surrounding drivers by exploiting a cloud computing system and VANET,
- **Sc.2 :** each vehicle can similarly transmit images for analyzing and mapping the road meteorology in real-time. Thus, drivers can define an itinerary based on meteorological criteria, notably to reduce the moving in areas having bad weather (e.g., snowy roads),
- **Sc.3 :** a vehicle can extract on-the-fly the license plate of preceding vehicles and then, sending the extracted plate characters to police services searching to localize stolen vehicles by matching extracted data with their reference databases,
- **Sc.4 :** similarly, a vehicle can extract on-the-fly people faces from the streets and then, sending the extracted face images to police services that aim to localize searched individuals.

In this study, we have experimented the proposed cloud-based system by considering the last scenarios related to the police service application.

2. Related Work

Nowadays, cloud computing developments are revolutionizing the world by providing to companies more and more powerful services. In particular, many companies tend to store their data on external servers or data centers. Indeed, this technology improves the Quality of Service (QoS) ; notably for the data management, the data security as well as for the data distribution. By this way, the providers of cloud computing systems allow many companies to develop services specifically focused on their principal activities. More precisely, cloud computing can be defined as a technology providing resources at three levels, namely infrastructures, software platforms and services [WSGB14]. The cloud computing was initially employed through wire-based network for internet and it has been progressively extended to the mobile network (e.g., through cellular networks). Notably, the cloud computing technologies facilitate the development of hybrid systems as well as the mutualizing of computational resources.

In this work, we are particularly interested by the development of cloud computing systems on the basis of VANET for enhancing and diversifying real-time road services. VANET networks have the particularity to exploit Ad-hoc systems. In other terms, these systems are self-organizing in the sense that each of them can communicate with others without the necessity of exploiting a pre-defined infrastructure. The development of VANET had a primary goal of supporting Intelligent Transport System through Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) communications (e.g., [Mas11]).

Besides, the novel generation of general public vehicles is equipped with computer-aided embedded navigation and vision systems such as Advanced Driver Assistance Systems (ADAS systems). In particular, ADAS systems are more and more employed for detecting road obstacles (e.g. ; self-parking) or for detecting the visibility degree of roads (e.g. ; automatic lighting systems). In parallel, experimental multi-camera vehicle systems are actively developed for the research in the fields of cartography and machine vision in order to reconstruct urban environments in 3D as well as to develop full autonomous navigation vehicles [HM13, HDS*13].

To the best of our knowledge, video services in vehicular clouds are not very developed. In [GWP13], Gerla et al. presented an image-on-demand service named “Pics-on-wheels” where some vehicles will send their acquired images for example by analyzing detected accidents. These images can then be used for assurance claims. In our case, we present a generic cloud computing system that could be used for developing various real-time video services by exploiting a distributed computing system. Notably, this system will be employed for sharing traffic information (e.g. ; in aided-navigation or road safety) by exploiting embedded vision-based systems (e.g., recognition system), CCSs and VANETs (see Figure 1).

3. Proposed Global Generic System for Real-time Road Video Scenarios

In our case, it is assumed that the vehicles will be equipped with embedded camera system, a GPS module and a VANET connecting system ($802.11p$). Notably, new generation vehicles are equipped with various types of sensors such as cameras located at the front and rear end. The proposed vision-based cloud computing system will take advantage of distributed computing and storing capabilities of conventional CCS and VANET (see Figure 1) for providing video services requiring high resources in term of data processing. In

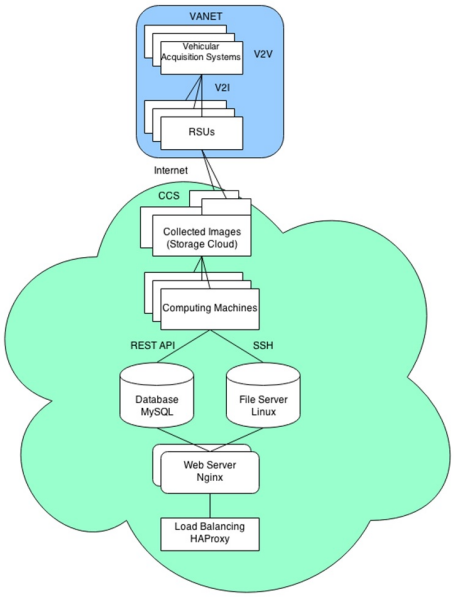


Figure 2: Proposed processing architecture of a global generic system.

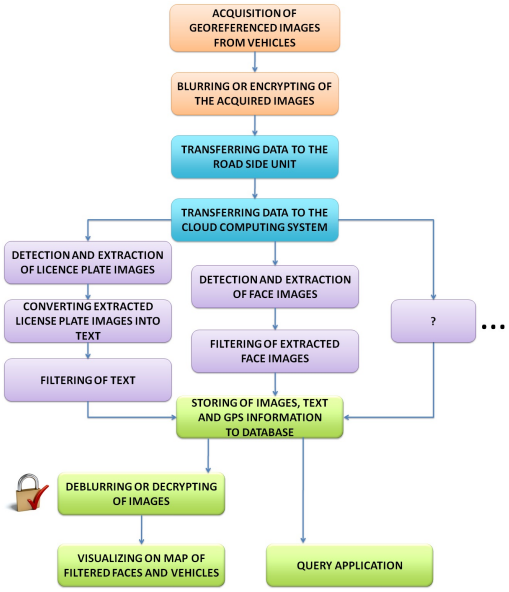


Figure 3: Proposed global dataflow diagram.

particular, the proposed system will be useful for visually recognizing dynamical objects of interest such as, for the search of stolen vehicles or individuals.

More precisely, the proposed system will exploit vehicular networks or external data center according to the needs. Yu et al. classify some cloud-based systems related to VANET [YZG*13]. First, vehicular cloud is exclusively composed of vehicles. It allows vehicles to dynamically schedule on demand computational and storage resources. Second, roadside cloud is composed of dedicated servers and RSUs (Road Side Units). The later permits access to the cloud. This cloud is exclusively used by vehicles localized within the radio coverage of the RSU. Vehicles roam between successive RSUs to continuously benefit from the service. Third, central cloud is based either on dedicated servers in the Internet or data centers on VANET itself. In our case, we are using the concept of Hybrid Vehicular Cloud (HVC) which shares the processing between the Vehicular Cloud (VC) and the central cloud.

Moreover, we visualize in Figure 2 the architecture that has been developed for supporting the various data transfer and data processing. First, vehicles communicate with internet access point by using vehicle to infrastructure (V2I) or vehicle to vehicle (V2V) communications. RSUs are exploited for removing redundancy in captured images and GPS information. Second, the collected georeferenced raw data are then sent to a customized storage cloud (e.g. ; Amazon cloud). Computing machines continuously run the face extraction, GPS extraction and number plate recognition algorithms in parallel. The extracted license plate numbers as well as the extracted GPS information are saved in a database (textual information). The extracted images are copied to file servers. Users access the service by connecting to a load balancing server, which distributes the requests to several working web servers.

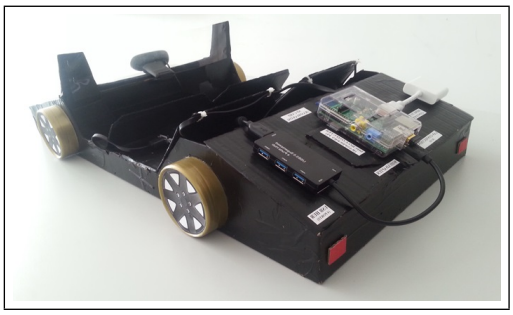
In Figure 3, we observe the global dataflow diagram of experimented scenarios (Sc.3 – 4). As can be observed, it worth mentioning that our architecture can also be used for the processing of other scenarios related to new real-time road video services (e.g. ; Sc.1 – 2).

4. Experimental Results

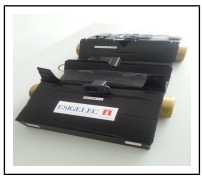
4.1. Developed Indoor Vehicular Monitoring Simulator

Figure 4 illustrated the developed embedded vehicular monitoring simulator. This vehicular monitoring simulator is composed of a car prototype (a rigid mock-up) as well as its associated vision-based embedded system (see Sub-figure 4a). This embedded system is equipped with a Logitech HD camera (see Sub-figure 4b) connected to a Raspberry Pi micro-computer (see Sub-figure 4c). This micro-computer includes a SD card for storing the acquired images. For simulating the moving of the car prototype, a screen has been placed in front of the webcam and a video corresponding to a vehicle path acquired by an external Mobile Mapping System has been filmed (e.g., videos from the Kitty research dataset † [GLU12, FKG13]). For such databases, the GPS information related to the images are provided. The micro-computer includes a wifi adapter that was used for simulating the VANET network. This embedded car prototype is connected to three workstations, the one simulating the RSU, the two others simulating the cloud nodes.

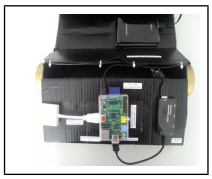
†. http://www.cvlibs.net/datasets/kitti/



(a) Global view of the car-based monitoring simulator.



(b) Acquisition part.



(c) Processing part.

Figure 4: Illustration of the developed car prototype and its associated vision-based embedded system (indoor vehicular monitoring simulator).

4.2. Implemented Architecture of the Proposed Global Generic System

More precisely, two python scripts are running on Raspberry Pi, one aims to capture images and to geo-tag them, and the other aims to transfer the images to RSU by FTP. In RSU, a bash script is written to send those images to two simulated cloud nodes by using SSH. By this way, the data flow is evenly distributed to the cloud nodes through WiFi. The computing machines (also cloud nodes) will process the images in storage servers and get the extracted faces, license plates and GPS information by running a python script invoking the corresponding algorithms. On the web server, we implemented a RESTful API to access the database. The extracted images are archived in file servers, while the license number, GPS and time are updated to the database. Thus, the updated information can be visualized. Moreover, new extraction algorithms can be developed for various query applications.

4.3. System Application and Evaluation

In this study, applications related to police services previously mentioned (Scenarios $3-4$) have been experimented by deploying computer vision approaches well-known for their efficiency on the proposed generic processing architecture (one simulated mobile node). Notably, open-source CSharp Emgu CV routines [‡] have been exploited for carrying out the face extraction as well as the OCR-based license plate extraction. Data matching has been experimented by comparing extracted features with a reference database generated by an operator. The proposed experimentation pipeline distributes the flow of collected images and extracted features are localized and labeled on Google Maps-based application in quasi real-time. Time information associated to the data transfer and data processing for one image of $16.5Kb$ (resolution of 640x480) can be observed in Table 1.

‡. http://www.emgu.com/

Image transferring	Time (sec.)	Image processing	Time (sec.)
Rasberry Pi to RSU	1.33	Face extraction	1.08
RSU to cloud nodes	1.12	License plate extraction	3.29

Table 1: Time information associated to the data transfer and data processing for one image. Data is processed on Intel Core i5 workstations of $2.4GHz$ under Windows 8.1 $64-bit$ with $4GB$ of RAM.

5. Conclusions and Future Works

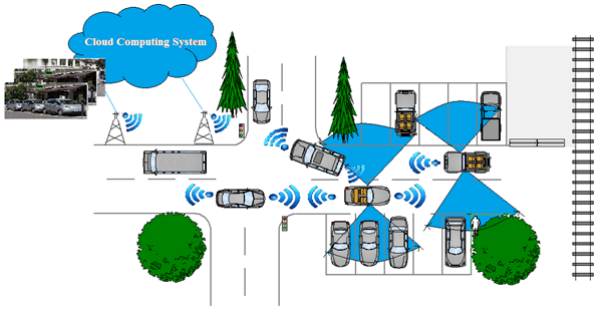


Figure 5: Illustration of a scenario related to the detection of available parking areas by exploiting VANET.

This paper presents our initial investigations for the design, the implementation and the simulation of a cloud computing system for enhancing and diversifying real-time video services through VANET and Onboard Navigation Systems. A vehicular monitoring simulator has been developed for carrying out indoor experiments. A generic hardware and software architecture is proposed for experimenting new video service applications.

Accordingly, next stage will consist of transferring this technology on two modular chassis that will be fixed on vehicle windshields for experiments in real mobile conditions (i.e., two moving nodes). Moreover, research will be pursued in indoor for improving the architecture of the developed simulator and simulations of the network architecture will be implemented under *ns*2 and *ns*3 Network Simulators [§] [¶]. Furthermore, we will tackle research in imagery for the detection of available parking areas in order to develop parking services. A corresponding targeted application was described in Scenario 1 and has been illustrated in Figure 5.

6. Acknowledgements

This work is part of the SAVEMORE project [‖]. The SAVEMORE project has been selected in the context of the INTERREG IVA France (Channel) - England European cross-border co-operation programme, which is co-financed by the ERDF.

§. http://nsnam.isi.edu/nsnam/
¶. http://www.nsnam.org/
‖. http://www.savemore-project.eu/

References

[FKG13] FRITSCH J., KUEHNL T., GEIGER A. : A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)* (2013).

[GLU12] GEIGER A., LENZ P., URTASUN R. : Are we ready for autonomous driving ? the KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[GWP13] GERLA M., WENG J. T., PAU G. : Pics-on-wheels : Photo surveillance in the vehicular cloud. In *IEEE International Conference on Computing, Networking and Communications (ICNC)* (2013), pp. 1123–1127.

[HDS*13] HAMMOUDI K., DORNAIKA F., SOHEILIAN B., VALLET B., MCDONALD J., PAPARODITIS N. : A synergistic approach for recovering occlusion-free textured 3D maps of urban facades from heterogeneous cartographic data. *International Journal of Advanced Robotic Systems. Vol. 10* (2013), 10p.

[HM13] HAMMOUDI K., MCDONALD J. : Design, implementation and simulation of an experimental multi-camera imaging system for terrestrial and multi-purpose mobile mapping platforms : A case study. *Applied Mechanics and Materials, Trans Tech Publications, Selected papers from the International Conference on Optimization of the Robots. Vol. 332* (2013), 139–144.

[Mas11] MASLEKAR N. : *Adaptative traffic signal control system based on inter-vehicular communication*. Ph.D. thesis, University of Rouen, Esigelec School of Engineering, 2011.

[WSGB14] WHAIDUZZAMAN M., SOOKHAK M., GANI A., BUYYA R. : A survey on vehicular cloud computing. *Journal of Network and Computer Applications. Vol. 40* (2014), 325–344.

[YZG*13] YU R., ZHANG Y., GJESSING S., XIA W., YANG K. : Toward cloud-based vehicular networks with efficient resource management. In *IEEE Network* (2013), pp. 48–55.

Design, Implementation and Simulation of a Cloud Computing System for Enhancing Real-time Video Services by using VANET and Onboard Navigation Systems

Karim HAMMOUDI[1,2] Nabil AJAM[1,2] Mohamed KASRAOUI[1,2] Fadi DORNAIKA[3,4]

Karan RADHAKRISHNAN[2,*] Karthik BANDI[2,*] Qing CAI[2,*] Sai LIU[2,*]

[1] IIS Laboratory, Research Institute for Embedded Electronics Systems, Rouen, France

[2] Department of ICT, ESIGELEC School of Engineering, Rouen, France (*MS Students)

[3] Department of Computer Science and Artificial Intelligence, University of the Basque Country, San Sebastian, Spain

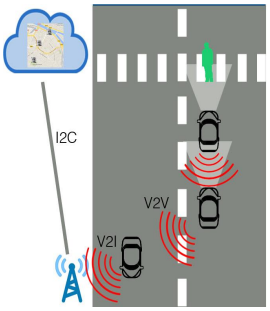
[4] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

I. ABSTRACT

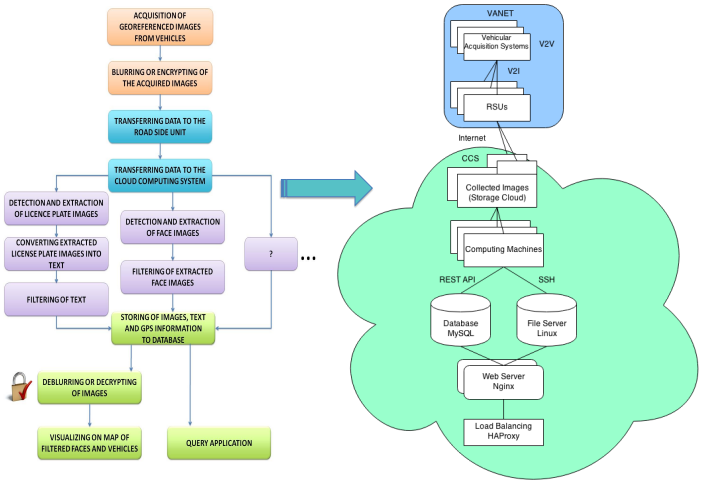
In this poster, we propose a design for novel and experimental cloud computing systems. The proposed system aims at enhancing computational, communicational and annalistic capabilities of road navigation services by merging several independent technologies, namely vision-based embedded navigation systems, prominent cloud computing systems and Vehicular Ad-hoc NETwork (VANET). This work presents our initial investigations by describing the design of the proposed system. The designed system has been simulated with various scenarios of video-based road services. Moreover, the associated architecture has been implemented on a small scale car prototype. The implemented architecture has been experimented in the case of a simulated road service to aid the police agency. The goal of this service is to recognize and track searched individuals and vehicles in a real-time monitoring system remotely connected to moving cars. The presented work demonstrates the potential of our system for efficiently enhancing and diversifying real-time video services in road environments.

II. INTRODUCTION AND MOTIVATION

- To exploit the cameras integrated on vehicles for proposing new road services (e.g. police or parking services)
- To extend the computational capabilities of embedded systems by using cloud computing systems
- To mutualize the collected road information for enriching monitoring systems
- To process acquired road images on-the-fly for real-time video services



III. PROPOSED CLOUD COMPUTING SYSTEM



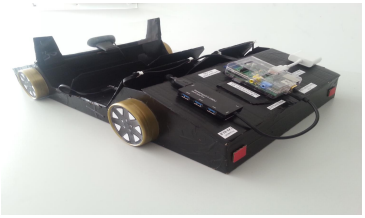
Global dataflow diagram

Cloud-based generic computing architecture

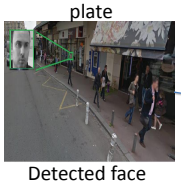
IV. SIMULATION AND EXPERIMENTAL RESULTS

Produced car prototype with associated embedded system

Detected license plate

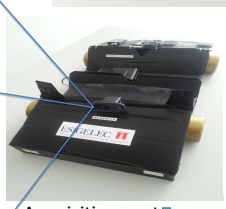


[1-3]



Detected face

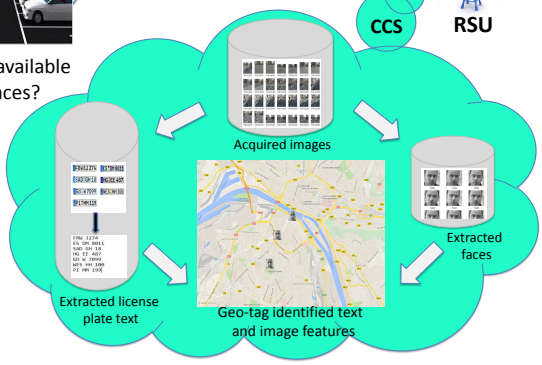
Detecting available parking places?



Acquisition part

Processing part

Road Images



V. CONTRIBUTIONS

- Design a hybrid computing system based on VANET and CCS for enhancing the road monitoring coverage
- Development of an indoor vehicular monitoring simulator for experimenting real-time video services
- Simulation of new service ideas for smart cities and Intelligent Transportation Systems

VI. FUTURE WORKS

- To improve the performance of the proposed architecture
- To simulate our architecture by using ns2 or ns3 Network Simulators
- To improve the efficiency of the recognition methods
- To transfer this embedded monitoring technology onto real vehicles
- To detect available parking places in collected road images
- To regulate the road traffic
- To study meteorological conditions for avoiding bad weather itineraries

VII. BIBLIOGRAPHY

[1] K. Hammoudi, F. Dornaika, B. Soheilian, B. Vallet, J. McDonald, N. Paparoditis. **A Synergistic Approach for Recovering Occlusion-free Textured 3D Maps of Urban Facades from Heterogeneous Cartographic Data**. In *International Journal of Advanced Robotic Systems*, Volume 10, 10p., **2013**.

[2] K. Hammoudi, J. McDonald. **Design, Implementation and Simulation of an Experimental Multi-camera Imaging System for Terrestrial and Multi-purpose Mobile Mapping Platforms: A Study Case**. In *Applied Mechanics and Materials* (journal), Volume 332, pp. 139-144, Trans Tech Publications, Switzerland, Selected papers from the International Conference on Optimization of the Robots, **2013**.

[3] K. Hammoudi, F. Dornaika, B. Soheilian, B. Vallet, J. McDonald, N. Paparoditis. **Recovering quasi-real occlusion-free textures for facade models by exploiting fusion of image and Laser street data and image inpainting**. In *Proc. IEEE International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Dublin, Ireland, May **2012**.

This work is part of the SAVEMORE project (http://www.savemore-project.eu).
The SAVEMORE project has been selected in the context of the INTERREG IVA France (Channel) – England
European cross-border co-operation programme, which is co-financed by the ERDF.

Vers un schéma temps réel de compression multi-vues sans perte

B. Battin[1] et J. Lehuraux[1] et P. Vautrot[2] et L. Lucas[2]

[1]OPEXMedia
[2]Université de Reims Champagne-Ardenne

Résumé

Cet article s'intéresse au problème de la compression multi-vues en environnement virtualisé. Nous présentons notamment un nouveau schéma de compression multi-vues sans perte destiné à des scènes virtuelles et basé sur l'algorithme LOCO-I. Notre algorithme exploite la double redondance (spatiale et temporelle) spécifique à ce type de média en adaptant les étapes de prédiction et de modélisation de contexte à la matrice d'images. Les premiers tests effectués avec notre approche montrent que celle-ci propose de bons ratios de compression pour une complexité algorithmique moindre vis-à-vis des méthodes de l'état de l'art.

This paper investigates the problem of multiview video compression in virtualized environments. In particular, we present a new lossless compression scheme adapted to virtual scenes and based on the LOCO-I algorithm. This algorithm exploits the double data redundancy inherent to such media by extending the prediction and the context modeling step along the view and time axes. The preliminary results show that our approach produces good compression ratios and outperforms state-of-the-art lossless methods on this kind of data.

Mots clé : Multi-vues, 3DTV, compression 3D, auto-stéréoscopie, compression sans perte

1. Introduction

Durant la dernière décennie, l'accélération spectaculaire des évolutions technologiques a fondamentalement changé notre relation aux médias. Dans le domaine du multimédia, de nouvelles technologies, telles que la télévision 3D (3DTV) ou la "Free Viewpoint Television" (FTV), disposent maintenant d'une qualité d'image ainsi que d'un confort visuel au moins comparable à la télévision 2D. Dans un contexte différent, la virtualisation d'applications professionnelles graphiques 3D n'était pas considérée jusqu'à maintenant comme viable compte tenu des limitations en terme de technologie et de coût. Mais grâce aux dernières avancées en terme d'accélération GPU, plusieurs acteurs ont commencé à déployer leurs applications graphiques hautes performances sur des systèmes virtualisés. Ces récents progrès soulèvent toutefois certaines questions : comment ces deux domaines peuvent-ils co-exister ? Comment assurer une expérience immersive de qualité pour l'utilisateur qui souhaite évoluer et interagir librement dans le monde virtuel 3D comme si celui-ci était physiquement présent ? Délivrer du contenu multi-vues en temps réel sur le réseau peut être très coûteux en terme de bande passante et de délais. Les protocoles de streaming multi-vues interactifs deviennent un challenge technique substanciel. En particulier, ils doivent trouver un compromis entre efficacité de codage et flexibilité de navigation afin de délivrer les images 3D dans de bonnes conditions (sans perte de qualité et avec une bonne restitution 3D). Les images (et vidéos) multi-vues sont généralement produites à l'aide d'un ensemble de n caméras synchronisées, réelles ou virtuelles, qui capturent une même scène depuis différents points de vue. Ce type de système génère une quantité de données conséquente (typiquement n fois la taille d'un flux vidéo standard), qui nécessite d'être compressée efficacement pour pouvoir être stockée ou transmise sur le réseau. Pour qu'un schéma de compression multi-vues soit considéré comme pertinent, celui-ci doit prendre en compte la double redondance spatiale et temporelle spécifique aux séquences multi-vues. Parmi les différentes approches possibles, on distingue deux familles majeures : la première utilise directement les données multi-vues comme H.264/MVC [CWU*09, MSMW07] ou le prochain standard 3D-HEVC [Ohm13,MJCPP13]. La deuxième approche consiste à essayer d'atteindre un codage optimal en convertissant les données multi-vues dans un format 3D spécifique (MVD, LDI [JMG09] ou DES [SMM*09]), puis en codant cette nouvelle donnée avec les outils adaptés. L'utilisation de ces formats 3D permet d'exploiter la corrélation spatiale entre les vues et de réduire de manière significative la quantité d'informations à coder en aval. Bien que le nombre d'algorithmes de compression multi-vues devienne

conséquent, la compression sans perte de tels média est rarement évoquée. Pourtant, la compression multi-vues sans perte peut etre considérée comme nécessaire dans certains domaines applicatifs tels que les données médicales ou les rushs cinéma. Dans ce papier, nous présentons notre schéma de compression multi-vues sans perte : Multi-LS. Notre méthode, basée sur l'algorithme LOCO-I [WSS96,WSS00], exploite la double redondance en adaptant les étapes de prédiction fixe et de modélisation de contexte de l'algorithme à la matrice d'image. Nous avons choisi de baser notre approche sur l'algorithme LOCO-I car celui-ci a l'avantage de proposer un coût algorithmique très faible et d'être facilement parallélisé sur le serveur afin de régler les problèmes de délais évoqués dans le précédent paragraphe. Afin d'évaluer les performances de notre algorithme, en terme de ratio de compression, nous comparons nos résultats à plusieurs standards de compression sans perte issus de l'état de l'art. Cet article est organisé de la manière suivante : dans la section 2, nous dressons l'état de l'art actuel des techniques de compression vidéo sans perte. La section 3 présente l'algorithme de base puis met en avant les principaux concepts de chaque étape. Dans la section 4, nous décrivons et validons l'adaptation des étapes de prédiction fixe et de modélisation de contexte dans le cadre des séquences multi-vues. La section 5 est, quant à elle, dédiée à la présentation des résultats obtenus par Multi-LS comparés aux méthodes de l'état de l'art. Nous concluons enfin cet article en section 6.

2. Etat de l'art

Avec l'émergence, ces dernières années, de la télévision 3D, on a pu observer l'apparition de nouveaux algorithmes de compression dédiés à ce nouveau type de média très coûteux en terme de stockage ou de débit. Parmis ceux-ci, on compte notamment le standard H.264/MVC ainsi que son récent successeur 3D-HEVC (basé sur H.265). Toutefois, on constate que la question de la compression multi-vues sans perte n'est pas abordée dans la littérature et les deux standards précédemment cités n'offrent pas la possibilité de compresser le signal d'entrée sans distortion. Dans certains secteurs d'activité (comme la production cinématographique ou l'industrie médicale), il est cependant préférable de conserver ou de transmettre les données sans apporter de distortion au signal d'entrée afin de faciliter le travail de post-production dans le cas du cinéma et d'éviter les erreurs d'interprétation sur les données patient dans le domaine médical.

La solution généralement adoptée consiste alors à utiliser des outils standards de compression vidéo sans perte pour encoder séparement chaque flux vidéo correspondant aux N vues de la séquence auto-stéréoscopique. Cette aproche n'est cependant pas considérée comme optimale d'un point de vue efficacité de compression car celle-ci permet uniquement de prendre en compte la redondance temporelle au sein des images appartenant au flux associé à une vue donnée, alors que la redondance spatiale présente entre chaque vue n'est pas exploitée.

Parmi les outils disponibles pour compresser un signal vidéo sans perte, on trouve tout d'abord le nouveau standard H.265/HEVC [SOHW12, OSS*12] (ou MPEG-H partie 2) qui permet d'obtenir de meilleurs taux de compression que son prédécesseur H.264 [WSBL03] (ou MPEG-4 AVC). Ce gain de performance en terme de compression est notamment dû à l'utilisation d'algorithmes similaires à H.264 mais plus complexes sans pour autant augmenter de manière réellement significative les temps d'encodage. On considère généralement qu'à qualité égale entre H.264 et H.265/HEVC, le débit de la vidéo H.265/HEVC est deux fois moindre que celle compressée à l'aide d'H.264. Ces deux algorithmes offrent la possibilité de compresser sans perte le signal d'entrée en désactivant certaines étapes du pipeline d'encodage (voir la figure 1) tout en gardant les mécanismes de base de la compression vidéo à savoir l'estimation/compensation de mouvement ainsi que l'intra-prediction. L'inconvénient majeur de ces deux algorithmes provient de ces deux mécanismes, et plus particulièrement de l'estimation/compensation de mouvement, qui requiert beaucoup de temps de calcul comme nous le constaterons dans la section 5.

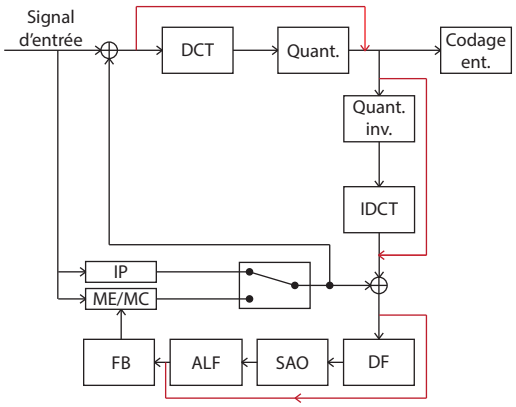


Figure 1: *Pipeline de codage sans perte pour HEVC (IP : Intra-Prediction, ME : Motion Estimation, MC : Motion Compensation, DF : Deblocking Filter, SAO : Sample Adaptive Offset, ALF : Adaptive Loop Filtering, FB : Frame Buffer).*

En marge de ces deux standards largement utilisés, on trouve plusieurs algorithmes fonctionnant tous de manière relativement similaire notamment HuffYUV [RG00], LOCO-I [WSS96] ou encore Lagarith [Gre04]. Ces trois algorithmes fonctionnent tous conformément au pipeline décrit par la figure 2 à savoir :

- Une étape de conversion colorimétrique, facultative suivant la prise en charge de l'algorithme, permettant une décorrélation préalable des données au niveau de l'espace de couleur.
- Une étape de prédiction qui va générer à partir de pixels voisins du pixel en cours de codage une valeur de prédiction à l'aide d'une fonction de prédiction spécifique. La différence entre la valeur de prédiction et la valeur du réel constitue alors l'erreur de prédiction.
- Enfin une étape de codage entropique de l'erreur de prédiction.

La table 1 permet d'identifier les différences notables entre ces trois algorithmes, relatives par exemple, à la fonc-

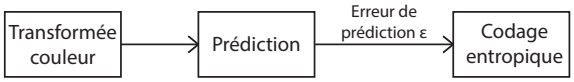


Figure 2: *Pipeline général des approches basées prédiction.*

	HuffYUV	Lagarith	LOCO-I
Espaces couleurs	RGB YUY2	RGB YUY2 YV12	RGB YUV
Prédiction	Gauche Gradient Médian	Médian	Médian
Codage entropique	Huffman	Arithmétique	Rice-Golomb
Inter prédiction	Non	NULL frames	Non
Run length	Non	Oui	Oui

Table 1: *Comparaison de codecs à base de prédicteurs*

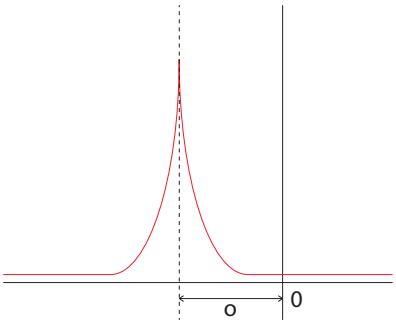


Figure 3: *Distribution des erreurs de prédiction fixe de la MED.*

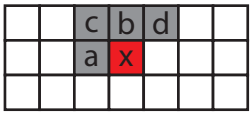


Figure 4: *Template $\mathcal{T}$ du pixel x.*

tion de prédiction utilisée, au type de codage entropique ou même à la présence ou non de mécanismes de prédiction inter-images. Dans ce dernier cas, on constate que Lagarith supporte également les “null frames”, ce qui lui permet d’éviter d’encoder l’image courante et d’utiliser l’image précédente lors du décodage dans le cas où ces dernières seraient mathématiquement identiques.

Toutes les approches précédemment citées dans cette section permettent d’exploiter la corrélation temporelle présente au sein de plusieurs frames consécutives et ainsi, de réduire de manière conséquente le volume initial de données. Dans le cas de la compression de vidéos multi-vues, l’algorithme doit aussi prendre en compte la corrélation spatiale entre les images issues de deux vues adjacentes au même temps t. Cette corrélation peut être réduite par des méthodes basées sur la compensation de disparité via l’utilisation de cartes de profondeur, comme c’est le cas pour le futur standard 3D-HEVC [MJCPP13]. Toutefois, ce type de processus rajoute un volume supplémentaire de données à compresser (sans perte) et peut être incompatible avec des applications de type rendu volumique pour lesquelles il est difficile de décider quelle profondeur utiliser pour un pixel donné. Pour ces deux raisons, nous avons opté pour une approche alternative qui sera présentée en section 4.

3. Les fondements de MULTI-LS

Multi-LS dérive de l’algorithme JPEG-LS qui constitue le standard ISO pour la compression d’image sans perte. Cette norme est basée sur l’algorithme LOCO-I (LOw COmplexity LOssless COmpression for Images) mis au point par [WSS96]. LOCO-I repose sur le concept de la modélisation de contexte et permet d’obtenir des ratios de compression similaires à l’algorithme CALIC [Wu95] tout en proposant une complexité algorithmique moindre. L’algorithme travaille de manière séquentielle sur chaque pixel de l’image, et pour chacun d’eux, procède aux cinq étapes suivantes.

3.1. Prédiction

Tout d’abord, l’algorithme doit générer une valeur de prédiction notée $\hat{x}_{MED}$ pour le pixel courant x à partir d’un voisinage constitué des pixels a, b, c et d (*cf.* figure 4). Cette opération est réalisée à l’aide d’une fonction de prédiction fixe appelée “median edge detector” (MED) et définie par l’équation 1. Celle-ci permet la détection d’éventuels bords (horizontaux ou verticaux) autour de x et ainsi d’adapter la valeur de prédiction en conséquence.

$$\hat{x}_{MED} = \begin{cases} \min(a,b) & \text{si} \quad c \geq \max(a,b) \\ \max(a,b) & \text{si} \quad c \leq \min(a,b) \\ a+b-c & \text{sinon} \end{cases} \tag{1}$$

A partir de $\hat{x}_{MED}$ et de x, on est maintenant en mesure de calculer l’erreur de prédiction fixe, notée ε_{MED}, qui est la différence entre $\hat{x}_{MED}$ et x. L’état de l’art, et tout particulièrement [NL80], nous indique que la distribution des erreurs de prédiction fixe générées par des fonctions de prédiction fixe (telle que MED) est modélisable par une distribution géométrique. Dans la section 3.3, nous verrons comment ce décalage est compensé grâce aux diverses informations associées au contexte courant afin d’assurer un codage entropique optimal.

3.2. Modélisation du contexte

Durant cette étape, l’algorithme va utiliser le pixel courant x ainsi que son template $\mathcal{T}$ afin de lui associer un contexte $\mathcal{C}_x$. Ce template est constitué des quatre pixels a, b, c et d appartenant au voisinage de x (*cf.* figure 4). L’algorithme calcule ensuite trois gradients à partir des valeurs des pixels appartenant à $\mathcal{T}$, conformément à l’équation 2.

$$\begin{aligned} g_1 &= d-b \\ g_2 &= b-c \\ g_3 &= c-a \end{aligned} \tag{2}$$

où la valeur de chaque gradient $g_i (1 \leq i \leq 3)$ est comprise dans l’intervalle $[-255;255]$.

Le triplet $\{g_1, g_2, g_3\}$ va nous permettre de caractériser l'activité autour du pixel x et de définir un modèle probabiliste pour le contexte C_x qui lui est associé. Dans [FWA04], l'auteur évoque un des problèmes récurrents dans les approches basées sur la modélisation de contexte : la dilution de contexte. Ce phénomène émerge généralement lorsqu'un trop grand nombre de contextes est utilisé, rendant insuffisante la quantité de statistiques pour chaque contexte et ayant un impact sur les performances en terme de compression. Afin de pallier cela, le nombre total de contextes est réduit par quantification des gradients g_i en q_i où $q_i \in [-4;4]$ et en fusionnant les contextes de signes opposés (l'information de signe est toutefois préservée dans une variable *SIGN* et utilisée dans la section 3.3). Ainsi, le nombre de contextes est réduit de 511^3 à $(9^3+1)/2 = 365$. Si $q_1 = q_2 = q_3 = 0$, l'algorithme considère que le pixel x appartient à une zone stationnaire de l'image. Dans ce cas, on comptabilise le nombre de pixels ayant la même valeur que x et cette chaîne est ensuite codée à l'aide de l'algorithme "Block-MELCODE" [OUO77].
Chaque contexte contient 4 compteurs *A, B, C* et *N* utilisés dans les deux prochaines sections :

- *A* accumule la valeur absolue des erreurs de prédiction (notée $|\varepsilon|$) et permet le calcul du paramètre *k* pour le codage entropique (*cf.* section 3.4) à l'aide du codeur de Rice-Golomb [Gol66]. *A* est initialisé à 4.
- *B* contient l'information nécessaire au calcul de *C* durant l'étape de correction adaptative (*cf.* section 3.3). Sa valeur initiale est 0.
- *C* contient la valeur de correction associée au contexte courant. Comme *B*, celui-ci est initialisé à 0.
- *N* contient le nombre d'occurences pour le contexte et est initialisé à 1.

3.3. Correction adaptative

Dans la section 3.1, nous avons évoqué la nécessité de compenser le décalage o entre le centre de la distribution géométrique associée au contexte courant C_x et l'origine. o est une valeur réelle pouvant s'exprimer sous la forme suivante : $o = R - s$ où R est entier et $s \in [0;1[$. L'étape de correction adaptative va stocker une approximation de R dans C et l'appliquer à la valeur $\hat{x}_{MED}$ précédemment calculée. Ainsi, l'amplitude des erreurs de prédiction corrigées ε se situe dans l'intervalle $]-1;0]$. Cette approximation peut être calculée de manière efficace en utilisant l'équation 3 où D représente la somme de toutes les erreurs de prédiction ε pour le contexte courant.

$$C = \lceil D/N \rceil \tag{3}$$

Toutefois, LOCO-I n'utilise pas cette manière de faire pour deux raisons majeures. Tout d'abord, l'équation 3 introduit une division qui n'est pas compatible avec une faible complexité algorithmique. Ensuite, la présence ponctuelle d'erreurs à forte amplitude affecterait de façon critique les valeurs suivantes de C impactant en aval le codage entropique. LOCO-I propose donc une méthode appelée "Adaptive bias cancellation" [WSS96] afin de contourner les deux problèmes précédemment evoqués : la valeur de l'erreur de prédiction corrigée peut être calculée à partir des compteurs *B* et *N* en utilisant l'équation 4.

$$\varepsilon = \hat{x}_{MED} + SIGN * C \tag{4}$$

Dans la section suivante, nous décrivons le processus de codage entropique pour la valeur de prédiction corrigée ε.

3.4. Codage entropique

L'étape de codage entropique de l'erreur de prédiction corrigée ε est réalisée grâce à un codage de Rice-Golomb. La littérature [GV75] s'accorde sur le fait que les codages de Rice-Golomb sont optimaux pour les distributions géométriques "one-sided" d'entiers non négatifs. En notant que la valeur de ε appartient à l'intervalle $[-255;255]$, nous devons dans un premier temps réduire ce dernier à $[0;255]$. La première opération consiste à fusionner l'extrêmité négative (respectivement extrêmité positive) de la TSGD ("Two-Sided Geometric Distribution") avec la partie centrale négative (respectivement positive), réduisant ainsi les valeurs de l'intervalle à $[-128;127]$. Cette manipulation n'affecte pas particulièrement la TSGD puisque le nombre d'occurences dans les parties extrêmes est négligeable vis-à-vis des parties centrales. Ensuite, l'intervalle $[-128;127]$ est porté à $[0;255]$ en entrelaçant les valeurs positives et négatives à partir de 0. Nous devons maintenant calculer le paramètre *k* qui régit la longueur du code binaire produit correspondant à ε. Suivant [GV75], une bonne estimation de la valeur optimale de *k* peut être obtenue à partir de l'équation 5 (E représente l'espérance mathématique).

$$k = \lceil \log_2 E\,[|\varepsilon|] \rceil \tag{5}$$

En utilisant les compteurs *A* et *N* de C_x, nous pouvons calculer $E\,[|\varepsilon|]$ par le quotient A/N. Ainsi, la valeur de *k* peut être déterminée par l'équation 6 et ε est codé entropiquement.

$$k = \min_{k'} \{k' | 2^{k'} N \geq A\} \tag{6}$$

3.5. Mise à jour du contexte

L'étape finale consiste à mettre à jour le contexte courant C_x pour prendre en compte les statistiques associées au pixel précédemment codé x. L'erreur de prédiction corrigée ε est ajoutée à *B* tandis que $|\varepsilon|$ est ajouté à *A* et *N* est incrémenté. LOCO-I effectue également une procédure de "réinitialisation" : si *N* est plus grand qu'un nombre prédéfini (entre 32 et 256), les compteurs *A, B,* et *N* sont divisés par 2, incrémentant de manière significative le poids des récentes statistiques pour ce contexte. Enfin, *C* est mis à jour par la réalisation de la procédure de compensation du biais.

4. Description de l'algorithme MULTI-LS

Dans cette section, nous présentons dans un premier temps notre adaptation de LOCO-I pour les séquences d'images multi-vues puis nous démontrons la pertinence d'une telle approche.

4.1. Adaptation aux séquences multi-vues

Les modifications principales apportées à l'algorithme LOCO-I résident dans les étapes de prédiction et de modélisation de contexte. Dans la section 3.1, nous avons pu

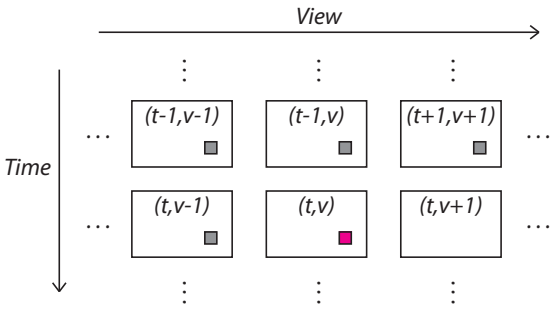


Figure 5: *Template Multi-LS.*

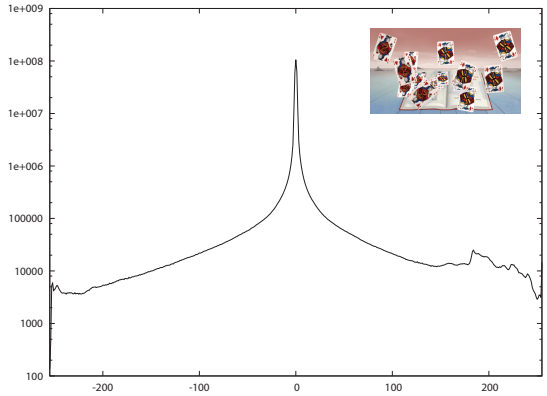


Figure 6: *Répartition des résidus de prédiction corrigés sur la séquence "Cartes".*

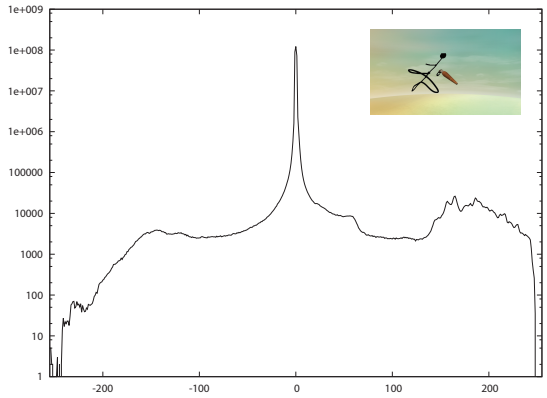


Figure 7: *Répartition des résidus de prédiction corrigés sur la séquence "Pinceau".*

voir que LOCO-I utilisait 4 pixels a, b, c et d (*cf.* figure 4) pour produire une valeur de prédiction fixe $\hat{x}_{MED}$ et pour associer un contexte $\mathcal{C}_x$ au pixel courant x. Notons (i, j) la position du pixel courant x (où i correspond à l'indice des lignes tandis que j correspond à l'indice des colonnes), les pixels $(i-1, j-1)$, $(i-1, j)$,$(i-1, j+1)$ et $(i, j-1)$ constituent le "template" $\mathcal{T}$ associé à x. Notre adaptation s'attache à exploiter le modèle probabiliste utilisé dans LOCO-I sur l'ensemble de la matrice d'image spécifique aux séquences multi-vues, plutôt que sur une seule image. De ce fait, le "template" $\mathcal{T}$ associé au pixel courant x (à la position (i, j)) sera constitué de pixels situés à la même position mais appartenant aux images voisines. Ainsi, notons $I_{t,v}(i, j)$ le pixel situé à la position (i, j) dans la vue v au temps t, l'étape de prédiction fixe considérera les trois pixels $a_m = I_{t,v-1}(i, j)$, $b_m = I_{t-1,v}(i, j)$ et $c_m = I_{t-1,v-1}(i, j)$. De la même façon, l'étape de modélisation du contexte considérera a_m, b_m, c_m et $d_m = I_{t-1,v+1}(i, j)$ en accord avec la figure 5. Tandis que LOCO-I exploite les redondances spatiales au sein d'une même image, notre schéma s'appuie sur les redondances inter-vues et temporelles : les gradients calculés permettent la caractérisation d'une activité temporelle (produite par un mouvement entre le temps t et le temps $t+1$) grâce à g_3 ainsi que d'une activité inter-vues (parallaxe entre la vue courante v et ses vues voisines $v-1$ et $v+1$) grâce à g_1 et g_3. Le mode "run-length" est toujours utilisé pour encoder les zones uniformes d'une vue. En effet, avant de calculer les gradients multi-vues, notre schéma considère les gradients locaux (ceux du standard LOCO-I). Si $g_1 = g_2 = g_3 = 0$, alors l'algorithme utilise le mode "run-length" pour encoder de manière efficace la zone uniforme, sinon, les pixels multi-vues sont pris en compte.
La prochaine section s'attache à confirmer la pertinence de notre adaptation multi-vues.

4.2. Validation du "template" multi-vues

L'efficacité de l'algorithme LOCO-I provient de son mécanisme de prédiction adaptatif (prédiction fixe et compensation adaptative du biais) permettant de produire des résidus de prédiction ε distribués selon une TSGD centrée en zéro. Cette distribution est ensuite entièrement exploitée par le codage Rice-Golomb pour produire des mots binaires de longueurs optimales. En conséquence, nous devons veiller à ce que notre adaptation ne perturbe pas ce comportement. La figure 6 de même que la figure 7 démontrent la distribution des résidus de prédiction ε pour deux de nos quatre ensembles de données multi-vues ("Cartes" et "Pinceau"). En analysant la figure 5, nous pouvons noter que la distribution des résidus de prédiction conserve le comportement attendu et l'utilisation du codage de Rice-Golomb pour l'étape de codage entropique est totalement justifié.

5. Résultats

Afin d'évaluer au mieux l'efficacité de notre méthode, nous proposons dans cette section de comparer nos résultats avec ceux obtenus à l'aide d'algorithmes de compression sans perte, issus de l'état de l'art. Pour cela, nous avons choisi d'inclure dans nos tests des méthodes de compression sans perte d'images (CALIC [Wu95] et JPEG-LS), des méthodes de compression vidéo sans perte (Huffyuv et Lagarith) ainsi que les standards actuels de compression video (HEVC et H.264) en utilisant un profil d'encodage sans perte. Nous ajoutons enfin une méthode de compression de données basée sur l'algorithme LZMA.

Les tests ont été réalisés sur 3 séquences de données multi-vues nommées "Cartes", "Pinceau" et "Rose", chacune constituée de 8 vues sur 25 images temporelles. Pour CALIC, JPEG-LS et LZMA, les tests ont été menés en com-

	Cartes	Pinceau	Rose
CALIC	0.39	0.67	0.45
JPEG-LS	0.38	0.66	0.45
LZMA	0.59	0.97	0.46
HEVC intra	0.58	0.77	0.58
HEVC	0.78	0.98	0.66
H.264 intra	0.58	0.76	0.55
Lagarith	0.56	0.75	0.58
Huffyuv	0.48	0.71	0.48
Multi-LS	**0.65**	**0.75**	**0.46**

Table 2: *Résultats obtenus (ratio de compression)*

	Temps d'encodage
CALIC	+
JPEG-LS	+ +
LZMA	+
HEVC intra	- -
HEVC	- - - -
H.264 intra	- -
Lagarith	+ + +
Huffyuv	+ +
Multi-LS	+ +

Table 3: *Comparaison subjective du temps moyen d'encodage pour chaque séquence*

pressant chaque image de manière indépendante, alors que pour les autres algorithmes (HEVC, H.264, Huffyuv, Lagarith) un encodage simulcast est réalisé sur chaque séquence. On peut noter la présence de deux profils d'encodage pour HEVC : un profil "intra" où chaque frame temporelle est une frame *I* et un profil plus standard, utilisant le mécanisme d'estimation/compensation de mouvement.

La table 2 présente les résultats obtenus en terme de ratio de compression et la table 3 propose une comparaison subjective des temps d'encodage pour chaque algorithme. La figure 8 reprend, quant à elle, les résultats précédemment obtenus afin de les présenter sous forme d'un nuage de points.

D'après la table 2, on constate logiquement que le profil standard de HEVC (avec estimation/compensation) de mouvement propose les meilleurs performances en terme de ratio de compression, alors que les profils intra HEVC et H.264 obtiennent des résultats relativement similaires. Toutefois, la table 3 nous indique que pour ces trois approches, le temps d'encodage moyen sur les différentes séquences est très élevé (770 secondes pour HEVC/intra, 4840 secondes pour HEVC et 900 secondes pour H.264). Rappelons que notre objectif était de proposer un algorithme de compression multi-vues sans perte temps réel. On ne peut donc pas considérer ces trois algorithmes comme viables dans le contexte qui est le notre.

On peut également remarquer, toujours d'après la table 2, que notre approche MULTI-LS permet d'obtenir des taux de compression équivalents ou supérieurs par rapport à Huffyuv, CALIC, JPEG-LS ainsi que Lagarith (sauf pour la séquence "rose" pour laquelle Lagarith est plus performant).

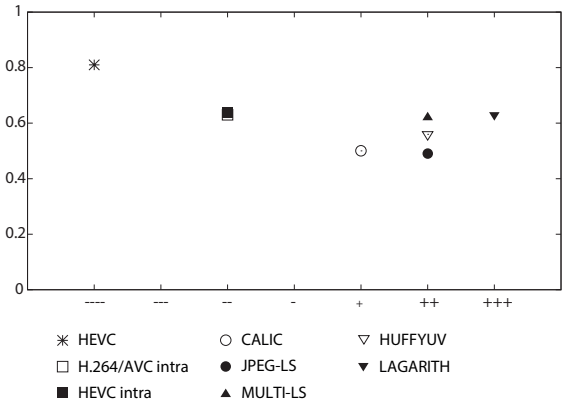


Figure 8: *Résultats obtenus.*

Au niveau des temps d'encodage, les algorithmes Huffyuv, JPEG-LS, Lagarith et notre approche Multi-LS permettent de compresser un flux vidéo monovue en temps réel. Cependant, il est important de noter que dans le cas de Huffyuv et de Lagarith, nous avons basé nos résultats sur des implémentations optimisées (SSE3, code ASM) alors que nous en sommes seulement à notre première implémentation en C de notre approche et qu'aucune optimisation n'a encore été réalisée (notamment une prochaine implémentation GPU sur CUDA). Nous espérons, par ce biais, proposer des temps d'encodage nettement inférieurs à ces standards, mais aussi permettre la compression de flux vidéos multi-vues en temps réel.

Enfin, concernant LZMA, on s'apperçoit que celui-ci propose de meilleurs taux de compression pour la séquence "Pinceau" uniquement. Cela peut être expliqué par le fait que la séquence "Pinceau" contient de larges zones uniformes où des méthodes basées sur un dictionnaire fournissent logiquement de meilleurs résultats.

6. Conclusion

Dans cet article, nous avons présenté un schéma de compression sans perte temps destiné à l'encodage temps réel de séquences multi-vues et basé sur LOCO-I. Les étapes de prédiction fixe et de modélisation de contexte sont adaptées à la matrice d'images multi-vues, exploitant ainsi les corrélations inter-vues et temporelle. Les résultats expérimentaux démontrent que le schéma proposé est capable de produire de meilleurs résultats que la plupart des algorithmes. Nous devons maintenant porter nos travaux sur la réalisation d'une implémentation GPU de Multi-LS, ce qui nous permettra de décharger le système dans les environnements virtualisés de sorte que les centres de traitement de données puissent proposer des expériences graphiques enrichies.

Références

[CWU*09] CHEN Y., WANG Y.-K., UGUR K., HANNUKSELA M., LAINEMA J., GABBOUJ M. : The emerging mvc standard for 3d video services. *EURASIP Journal on Advances in Signal Processing* (2009).

[FWA04] FORCHHAMMER S., WU X., ANDERSEN J. D. : Optimal context quantization in lossless compression of image data sequences. *IEEE Transactions on Image Processing* (2004).

[Gol66] GOLOMB S. W. : Run-length encodings. *IEEE Transactions on Information Theory. Vol. 12* (August 1966), 399–401.

[Gre04] GREENWOOD B. : Lagarith Lossless Video Codec. `http://lags.leetcode.net/codec.html`, 2004.

[GV75] GALLAGER R., VOORHIS D. V. : Optimal source codes for geometrically distributed integer alphabets. *IEEE Transactions on Information Theory* (1975).

[JMG09] JANTET V., MORIN L., GUILLEMOT C. : Incremental-ldi for multi-view coding. In *3DTV Conference : The True Vision - Capture, Transmission and Display of 3D Video* (2009).

[MJCPP13] MORA E., JUNG J., CAGNAZZO M., PESQUET-POPESCU B. : Initialization, limitation and predictive coding of the depth and texture quadtree in 3d-hevc video coding. *IEEE Transactions on Circuits and Systems for Video Technology* (2013).

[MSMW07] MERKLE P., SMOLIC A., MUELLER K., WIEGAND T. : Efficient prediction structures for multiview video coding. *IEEE Transactions on Circuits and Systems for Video Technology* (2007).

[NL80] NETRAVALI A., LIMB J. O. : Picture coding : A review. In *Proceedings of the IEEE* (1980).

[Ohm13] OHM J.-R. : Overview of 3d video coding standardization. In *Proceedings of 3DSA2013, Keynote speech 2* (2013).

[OSS*12] OHM J.-R., SULLIVAN G. J., SCHWARZ H., KAN T. K., WIEGAND T. : Comparison of the coding efficiency of video coding standards - including high efficiency video coding (hevc). *IEEE Transactions on Circuits and Systems for Video Technology. Vol. 22*, Num. 12 (2012), 1669–1684.

[OUO77] OHNISHI R., UENO Y., ONO F. : The efficient coding scheme for binary sources. *IECE of Japan* (1977).

[RG00] RUDIAK-GOULD B. : HuffYUV Lossless Codec. `http://neuron2.net/www.math.berkeley.edu/benrg/huffyuv.html`, 2000.

[SMM*09] SMOLIC A., MUELLER K., MERKLE P., KAUFF P., WIEGAND T. : An overview of available and emerging 3d video formats and depth enhanced stereo as efficient generic solution. In *Picture Coding Symposium* (2009).

[SOHW12] SULLIVAN G. J., OHM J.-R., HAN W.-J., WIEGAND T. : Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and Systems for Video Technology. Vol. 22*, Num. 12 (2012), 1649–1668.

[WSBL03] WIEGAND T., SULLIVAN G. J., BJONTEGAARD G., LUTHRA A. : Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology. Vol. 13*, Num. 7 (2003), 560–576.

[WSS96] WEINBERGER M. J., SEROUSSI G., SAPIRO G. : Loco-i : A low complexity, context-based, lossless image compression algorithm. In *Data Compression Conference* (1996).

[WSS00] WEINBERGER M. J., SEROUSSI G., SAPIRO G. : The loco-i lossless image compression algorithm : principles and standardization into jpeg-ls. *IEEE Transactions on Image Processing. Vol. 9* (August 2000), 1309–1324.

[Wu95] WU X. : Context selection and quantization for lossless image coding. In *DCC 95, Data Compression Conference* (1995).

Méthode d'optimisation pour l'appariement de pixels d'images stéréoscopiques basée sur une métrique conjointe entropie-distorsion

A. Kadaikar, A. Mokraoui et G. Dauphin

L2TI, Institut Galilée, Université Paris 13 Sorbonne Paris Cité
99, avenue Jean-Baptiste Clément 93430 Villetaneuse, France

Résumé
Cet article s'intéresse au problème de la mise en correspondance de pixels d'images stéréoscopiques pour estimer la meilleure carte de disparité au sens du critère entropie-distorsion. Dans la majorité des cas, les correspondants sont choisis au sens de la minimisation de l'erreur quadratique moyenne retenue comme critère de distorsion. Cependant pour l'appariement d'un même pixel, il est possible que plusieurs disparités soient candidates puisqu'elles répondent uniquement au critère de distorsion minimale. En revanche le choix adopté pourrait ne pas être en adéquation avec la réduction du coût de codage. Pour y remédier, cet article propose une approche d'optimisation où la métrique habituelle est remplacée par une métrique entropie-distorsion de façon à ce que les disparités sélectionnées réduisent non seulement la distorsion de l'image reconstruite mais également l'entropie associée à la carte de disparité. L'estimation de la carte de disparité s'appuie sur la construction séquentielle d'un arbre afin d'éviter une recherche exhaustive tout en assurant de bonnes performances en termes d'entropie-distorsion. A une profondeur donnée dans l'arbre, les M meilleurs chemins retenus selon le critère entropie-distorsion sont ensuite prolongés pour construire de nouveaux chemins. Ces chemins sont triés selon la métrique entropie-distorsion pour n'en retenir que les M meilleurs. Le processus est itéré jusqu'à la lecture du dernier pixel à apparier. Les résultats de simulation montrent que notre algorithme fournit de meilleurs résultats en termes d'entropie-distorsion comparé à la programmation dynamique.

This paper focuses on the matching problem of pixels in stereoscopic images to estimate the best disparity map under entropy-distortion criterion. In most cases, matches are selected according to the minimum mean square error used as a distortion criterion. However for matching the same pixel, it is possible that several disparities are candidates since they meet only the minimum distortion criterion. In contrast, the choice may not be adopted in adequation with the coding cost reduction. To address this, this paper proposes an optimization approach where the usual metric is replaced by a entropy-distortion metric so that the selected disparities not only reduce the distortion of the reconstructed image, but also the entropy associated with the disparity map. The estimate of the disparity map is based on the sequential construction of a tree to avoid an exhaustive search while ensuring good performance in terms of entropy-distortion. At a given depth in the tree, the M-best paths chosen by entropy-distortion criterion are then extended to build new paths. These paths are sorted according to the entropy-distortion metric to retain only the M-best. The process is iterated until reading the last pixel to be matched. Simulation results show that our algorithm provides better results in terms of entropy-distortion compared to dynamic programming.

Mots clé : Image stéréoscopique, image 3D, mise en correspondance, algorithme à M-chemins, optimisation, entropie, distorsion.

1. Introduction

Une image stéréoscopique (ou image 3D) permet de recréer une impression de profondeur dans la scène observée. Elle est composée de deux vues à savoir la vue droite et la vue gauche. Ces vues correspondent à la même scène et sont capturées à partir de deux points de vue légèrement différents. De ce fait, les images stéréoscopiques requièrent deux fois plus d'informations qu'une image 2D. Il est donc important de considérer la question du codage des images stéréoscopiques dans le cadre applicatif de stockage ou de transmission notamment pour des flux vidéos stéréoscopiques. Dans l'objectif de réduire la redondance inter-vue, certains

travaux proposent de reconstruire l'une des deux vues en exploitant la deuxième vue (considérée comme vue de référence) combinée à la carte de disparité associée à la vue à reconstruire. La carte de disparité s'avère en effet moins coûteuse en termes binaire qu'une des deux vues. C'est le processus d'appariement de pixels homologues dans les deux vues qui permet d'estimer la carte de disparité associée à l'une des deux vues. La qualité de la vue reconstruite en dépend fortement. Les méthodes développées doivent en effet faire face aux éventuelles changements de luminosité, aux zones texturées et également aux occultations.

Un état de l'art sur l'appariement d'images stéréoscopiques montre que plusieurs travaux ont déjà considéré la question de l'estimation des cartes de disparité. Le problème de l'appariement stéréoscopique est généralement formulé par le problème de minimisation d'une fonction d'énergie (coût global pour une approche globale) ou plusieurs fonctions d'énergie (coûts locaux pour une approche locale) [SS02, BBH03]. Les méthodes développées se différencient de manière générale par : (i) les primitives adoptées (par exemple pixels, points d'intérêts, segments, blocs, régions, bords) et leurs attributs (par exemple niveau de gris, contraste, composantes de couleur, position du segment, orientation du segment) ; (ii) le coût global de l'appariement (incluant également les coûts locaux d'appariement qui mesurent le degré de dissimilarité entre les deux primitives correspondantes) et le coût des contraintes (par exemple l'unicité, l'ordonnancement, le lissage) ; (iii) la taille de la fenêtre d'appariement ; (iv) la zone dagrégation (ensemble des pixels pour le calcul du coût de mise en correspondance) ; et (v) la méthode d'optimisation. L'objectif principal des méthodes d'optimisation est de minimiser le coût global ou local pour garantir le meilleur appariement des primitives. La recherche exhaustive (greedy search) du meilleur correspondant n'est pas intéressante en raison de sa charge de calcul trop élevée. La programmation dynamique a été l'une des premières méthodes d'optimisation a être exploitée en stéréo. Différentes versions ont été proposées. Dans [OK85], des contraintes de régularité ont été introduites pour optimiser les appariements notamment lorsque le lecture est réalisée selon un balayage en ligne. Dans [Vek05], Veksler a imposé des contraintes de lissage selon les deux directions horizontale et verticale dans le but de récupérer la vraie carte de disparité. D'autres méthodes d'optimisation ont également été exploitées, telles que la relaxation [Nas92], les graph cut [BVZ01, BG05], et la propagation de croyance [Sun03, TCC07].

Les travaux présentés dans cet article visent à développer un algorithme d'appariement pixel à pixel. La métrique de mise en correspondance s'appuie généralement soit sur l'erreur quadratique moyenne (EQM), soit sur l'erreur en valeur absolue moyenne (EAM). Notons que, pour un pixel donné, il est parfois possible d'obtenir un ensemble de candidats qui satisfait au critère imposé. Néanmoins, certains de ces candidats risquent de coûter plus chers que d'autres en termes de débit binaire. Pour résoudre cette question, nous proposons de remplacer la métrique traditionnelle (EQM, EAM ...) par une métrique conjointe entropie-distorsion. De ce fait, les disparités sélectionnées améliorent non seulement la qualité de l'image reconstruite mais également réduisent l'entropie de la carte de disparité. Ce problème est formalisé par une minimisation Lagrangienne où la fonction coût est choisie comme nouvelle métrique de mise en correspondance. Pour éviter une trop grande charge de calcul (recherche exhaustive), l'estimation de notre carte de disparité repose sur la construction séquentielle d'un arbre à M-chemins (c.a.d a breadth-first search algorithm). A chaque profondeur de l'arbre, l'algorithme prolonge les M meilleurs chemins jusqu'à l'appariement du dernier pixel.

Notre article est organisé comme suit. La section 2, après avoir introduit quelques notations, formalise notre problème d'optimisation. Notre algorithme d'appariement basé sur la métrique conjointe entropie-distorsion est ensuite développé. Les résultats de simulation sont ensuite discutés en section 3. La section 4 conclut notre article.

2. Algorithme proposé de mise en correspondance d'image stéréoscopique

L'objectif de notre algorithme est d'estimer la carte de disparité relative à la vue droite à partir de la vue gauche au sens du critère entropie-distorsion. La vue gauche est considérée ici comme vue de référence.

2.1. Formulation du problème d'appariement au sens du critère entropie-distorsion

Introduisons tout d'abord quelques notations avant de décrire notre algorithme. Supposons que les images sont rectifiées. I_g représente l'image de la vue gauche et I_d celle de la vue droite, toutes deux de taille $K \times L$. $I_d(i,j)$ (respectivement $I_g(i,j)$) correspond à l'intensité du pixel situé à la position (i,j) dans I_d (respectivement I_g). Notons $d(i,j)$ la disparité du pixel $I_d(i,j)$; et $\mathbf{d} = \{d(i,j) \text{ avec } i = 0,...,K-1;\ j = 0,...,L-1\}$ la carte de disparité qui minimise le coût global choisit comme étant l'erreur quadratique moyenne entre l'image originale associée à la vue droite et sa version reconstruite donnée comme suit :

$$E_{global}(\mathbf{d}) = \sum_{i=0}^{K-1}\sum_{j=0}^{L-1}(\widehat{I}_d(i,j) - I_d(i,j))^2$$
$$\text{avec } \widehat{I}_d(i,j) = I_g(i, j + d(i,j)), \qquad (1)$$

sous la contrainte de minimiser aussi l'entropie $H(\mathbf{d})$ de la carte de disparité. Ce problème est régit par le formalisme de Lagrange donné par l'équation suivante :

$$\widehat{\mathbf{d}} = \text{argmin} J(\lambda, \mathbf{d}) = \text{argmin}(E_{global}(\mathbf{d}) + \lambda H(\mathbf{d})), \qquad (2)$$

où λ représente le multiplicateur de Lagrange.

2.2. Algorithme d'optimisation sous-optimale

Pour résoudre l'équation (2), avec le souci non seulement de réduire la complexité de calcul mais également d'assurer une bonne estimation de la carte de disparité, nous nous sommes appuyés sur le principe de l'algorithme de décodage séquentiel à M-chemins exploité dans le domaine des communications pour estimer le flux transmis à travers un canal bruité selon un critère de vraisemblance [Jel69].

Pour limiter l'espace de recherche, sans trop pénaliser les

performances de notre algorithme, nous avons introduit une fenêtre d'appariement notée $W = [w_{min}, ..., w_{max}]$ de taille N. Elle est positionnée sur chaque pixel à apparier. L'image de référence est lue en parcourant ses lignes de gauche à droite.

L'algorithme d'optimisation proposé s'appuie sur la construction séquentiel d'un arbre. A chaque profondeur t de l'arbre, un pixel à apparier y est associé. La profondeur t dépend de la ligne courante i, de la colonne courante j et du nombre de colonne L de l'image I_d :

$$t = i \times L + j \text{ avec } i = 0, ..., K-1 \text{ et } j = 0, ..., L-1. \quad (3)$$

Supposons qu'à la profondeur $t-1$, les M meilleurs chemins aient été retenus. A chaque chemin retenu est associé un coût $J^k_{t-1}(\lambda, d)$ tel que :

$$J^k_{t-1}(\lambda, d) = E^k_{t-1} + \lambda H^k_{t-1} \text{ avec } k = 1, ..., M, \quad (4)$$

où E^k_{t-1} est la distorsion cumulée sur le chemin k jusqu'à la profondeur $t-1$; et H^k_{t-1} est une estimation de l'entropie relative aux choix des disparités effectués sur le chemin k jusqu'à la profondeur $t-1$. Notons qu'aux M meilleurs chemins sont associés M cartes de disparité notées S^k :

$$S^k = \{d^k_1, d^k_2, ..., d^k_{t-1}\} \text{ avec } k = 1, ..., M. \quad (5)$$

A la profondeur suivante (c.a.d. t), chacun des M-chemins retenus est prolongé par N nouvelles branches auxquelles sont associées une disparité w et une distorsion locale $E_{b^w_t}$ donnée par :

$$E_{b^w_t} = ((I_d(i,j) - I_g(i, j+w))^2 \text{ avec } w = w_{min}, ..., w_{max}. \quad (6)$$

Les distorsions cumulées des $M \times N$ chemins prolongés sont alors mises à jour comme suit :

$$E^m_t = E^k_{t-1} + E_{b^w_t} \text{ avec } m = 1, .., M \times N \quad k = 1, ..., M \text{ et } w = w_{min}, ..., w_{max}. \quad (7)$$

Notons que nous ne disposons pas de la distribution des probabilités des disparités, information nécessaire pour le calcul de l'entropie H^k_t à la profondeur t

$$H^k_t = -\sum_{w=w_{min}}^{w_{max}} \widehat{p}^k_t(d = w) log_2(\widehat{p}^k_t(d = w)) \quad \text{avec } k = 1, ...M \times N. \quad (8)$$

Pour y remédier, nous proposons une estimation des probabilités par une loi de mélange donnée par l'équation suivante :

$$\widehat{p}^k_t(d = w|d^k_1, d^k_2, ..., d^k_t) = C_a \times p_a(d = w) + C_{exp} \times p^k_{exp}(d = w|d^k_1, d^k_2, ..., d^k_{t-1}) + C_c \times p_c(d = w|d = d^k_t), \quad (9)$$

où les coefficients C_a, C_{exp} et C_c vérifient la relation suivante :

$$C_a + C_{exp} + C_c = 1. \quad (10)$$

$p_a(d = w)$ représente la probabilité *a priori* associée à la disparité w choisie parmi les disparités possibles définies dans la fenêtre de recherche à la profondeur t. Nous la représentons par une loi discrète uniforme :

$$p_a(d = w) = \frac{1}{N} \text{ avec } w = w_{min}, ..., w_{max}. \quad (11)$$

$p^k_{exp}(d = w|d^k_1, d^k_2, ..., d^k_{t-1})$ est déduite des disparités retenues jusqu'à la profondeur $(t-1)$ sur le chemin k (c.a.d. $d^k_1, d^k_2, ..., d^k_{t-1}$). Enfin $p_c(d = w|d = d^k_t)$ est la probabilité relative à la disparité choisie à la profondeur t. Elle est donnée par l'expression suivante :

$$p_c(d = w|d^k_t = w_c) = \begin{cases} 1 \text{ if } w = w_c \\ 0 \text{ if } w \neq w_c \end{cases} \quad (12)$$

En ce qui concerne les coefficients de l'équation (10), nous proposons de les paramétrer comme suit :

$$C_a = \frac{\beta a}{\beta a + b + c}; \; C_{exp} = \frac{b}{\beta a + b + c}; \; C_c = \frac{c}{\beta a + b + c};$$

avec

$$a = K \times L - t; \; b = t \text{ et } c = 1. \quad (13)$$

β est un paramètre inférieur à 1 que nous fixons de façon à pouvoir contrôler l'importance attribuée aux probabilités *a priori* lors de l'estimation de notre carte de disparité. Notons que les paramètres "a" et "b" évoluent au fur et à mesure que notre algorithme progresse dans son traitement. "a" est choisi de façon à décroître alors que "b" augmente durant le traitement. Ceci se justifie par le fait qu'au démarrage l'algorithme s'appuie davantage sur p_a par manque d'informations. De ce fait l'influence de p_a s'estompe au détriment de la probabilité p^k_{exp} qui prend le relai.

Les coûts J^k_t (voir équation (4)) sont calculés puis classés selon un ordre croissant. L'algorithme retient alors parmi les $M \times N$ nouveaux chemins les M meilleurs. Les cartes de disparité respectives (c.a.d S^k) sont mises à jour. L'algorithme itère ce processus jusqu'à l'appariement du dernier pixel de l'image. Le premier chemin correspondra alors à la meilleure carte de disparité en termes de compromis entropie-distorsion.

La Figure 1 donne un aperçu rapide de notre algorithme d'appariement avec $M = 2$ et $N = 5$. Pour plus de détails, les différentes étapes sont résumées ci-dessous :

Algorithme à M-chemins basé sur un critère conjoint d'entropie-distorsion pour l'appariement stéréoscopique.
Input : Image gauche I_g et image droite I_d de taille $K \times L$
Output : Carte de disparité dense estimée, relative à I_d ;
1. Initialisation des valeurs : λ ; M ; w_{min} ; w_{max} ; β ; $i = -1$ et $j = -1$;
2. Incrémenter de 1 le numéro de la ligne i ;
3. Incrémenter l'index j ;
4. Positionner la fenêtre sur le pixel $I_g(i, j)$;
5. Prolonger les M meilleurs chemins courants jusqu'à la profondeur t ;
6. Calculer les distorsions des $M \times N$ branches ;
7. Mettre à jour les distorsions des chemins prolongés ;
8. Estimer les probabilités de choisir les disparités pour chacun des chemins prolongés ;
9. Déduire l'entropie associée aux disparités de chaque chemin ;
10. Calculer le coût J^k_t pour chacun des chemins ;
11. Classer les chemins dans l'ordre croissant en fonction de leurs coûts respectifs ;
12. Sélectionner les M meilleurs chemins parmi les $M \times N$ chemins ;
13. Mettre à jour les M cartes de disparités des chemins retenus ;

14. Recommencer depuis l'étape 3 si $j < L$ sinon continuer ;
15. Recommencer depuis l'étape 2 si $i < K$ sinon continuer ;
16. Choisir la meilleure carte de disparité associée à I_d.

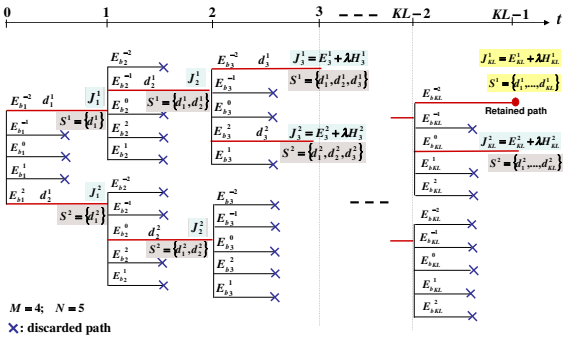


Figure 1: *Principe de notre algorithme d'appariement stéréoscopique avec $M = 2$ et $N = 5$.*

3. Analyse et discussion des résultats de simulation

Cette section présente les résultats de simulation fournis par notre algorithme d'optimisation pour la mise en correspondance d'images stéréoscopiques. Nos résultats sont comparés à ceux obtenus par l'algorithme de programmation dynamique de la toolbox "computer vision system" de Matlab [ref].

3.1. Image stéréoscopique de test : 'Tsukuba'

Les tests de simulation ont été réalisés sur les vues gauche et droite de l'image stéréoscopique "Tsukuba" de la base Middlebury [Mid]. La résolution spatiale de chaque image est de 288×384 pixels. La taille de la fenêtre d'appariement est fixée à 30 (c.a.d $N = 30$ avec $w_{min} = -15$ et $w_{max} = 14$) pour les deux algorithmes de mise en correspondance de pixels. La vue gauche est choisie comme vue de référence. La reconstruction de l'image droite repose sur la carte de disparité estimée soit par notre algorithme soit par l'algorithme de programmation dynamique. La Figure 2 montre l'image originale de la vue droite que nous cherchons à reconstruire à travers sa composante de luminance.

Figure 2: *Image originale de la vue droite.*

La Figure 3 est l'image reconstruite avec la carte de disparité estimée par la programmation dynamique basée sur des blocs de taille 1×1. Cette taille de bloc a été retenue parce que l'image reconstruite présente un meilleur *PSNR* à savoir 38.77 dB comparé aux autres tailles de blocs. La carte de disparité estimée (voir Figure 5) a une entropie égale à 3.65 bpd (bits par disparités).

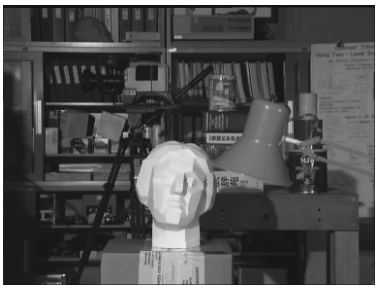


Figure 3: *Image reconstruite : carte de disparité estimée par l'algorithme de programmation dynamique (PSNR=38.77 dB ;H=3.65 bpd).*

Figure 4: *Image reconstruite : carte de disparité estimée par l'algorithme à M-chemins (PSNR=38.12 dB ; H=1.92 bpd).*

La Figure 4 présente l'image reconstruite avec la carte de disparité estimée par notre algorithme à M-chemins avec les paramètres suivants : $M = 4$, $\lambda = 1100000$ et $\beta = 0.2$. La qualité de la reconstruction évaluée en termes de *PSNR* est de 38,12 dB. Ce *PSNR* est équivalent à celui de la programmation dynamique. En revanche notre carte de disparité a une entropie égale à 1.92 bpd (voir Figure 6). Pour une qualité équivalente, notre algorithme réduit le débit de 48%. Ceci est confirmée par l'analyse des histogrammes relatifs aux deux cartes de disparité fournis par les Figures 7 et 8.

Les courbes de la Figure 9 illustrent les performances débit-distorsion des deux algorithmes d'appariement stéréoscopique. La courbe marquée par $'+'$, a été obtenue en jouant sur le paramètre λ. La courbe marqué par $'*'$ a été réalisée en agissant sur la taille des blocs (de 1×1 jusqu'à 17×17). Nous observons clairement l'avantage de notre algorithme par rapport à l'algorithme de programmation dynamique. En effet, pour un débit par exemple de l'ordre de 2.32 bpd nous obtenons un gain de 19.8 dB en faveur de notre méthode d'optimisation.

3.2. Image stéréoscopique de test : 'Journal'

D'autres tests ont été réalisés sur l'image stéréoscopique 'Journal' de la base Middlebury [Mid]. La résolution spatiale des images des vues droite et gauche est de 383×435 pixels. Une fenêtre centrée de taille $N = 30$ ($w_{min} = -15$, $w_{max} = 14$) est choisie pour les deux algorithmes d'appariement. L'image originale de la vue droite à reconstruire est donnée par la Figure 10.

La Figure 11 présente la meilleure reconstruction possible

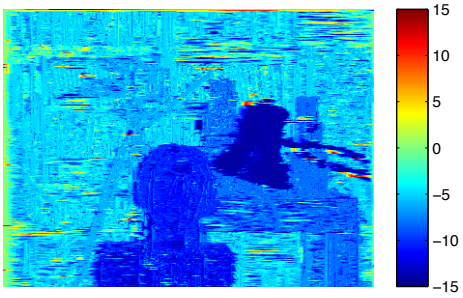


Figure 5: *Carte de disparité estimée par la programmation dynamique (PSNR=38.77 dB ;H=3.65 bpd).*

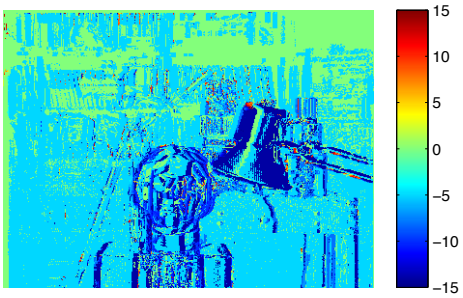


Figure 6: *Carte de disparité estimée par l'algorithme à M-chemins (PSNR=38.12 dB ; H=1.92 bpd).*

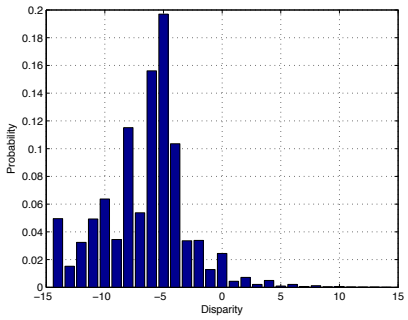


Figure 7: *Histogramme des disparités estimées par la programmation dynamique (PSNR=38.77 dB ;H=3.65 bpd).*

lorsque l'algorithme de programmation dynamique (avec des blocs de taille 1×1) est exploité pour estimer la carte de disparité. Le *PSNR* relatif à cette reconstruction est de 27.06 dB avec une carte de disparité estimée ayant une entropie égale à 4.46 bpd (voir Figure 13).

La Figure 12 correspond à l'image droite reconstruite avec la carte de disparité estimée (voir Figure 14) par notre algorithme à *M*-chemins (avec *M*=2, λ=10000000 et β=0.2). Le *PSNR* relatif à cette reconstruction est de 27.39 dB. L'entropie de notre carte de disparité est de 1.85 bpd. Pour un *PSNR* équivalent, nous obtenons un gain en débit de 58.5% par rapport à la programmation dynamique. Les histogrammes des cartes de disparité estimées confirment les avantages de notre méthode d'optimisation (voir Figure 15 et Figure 16).

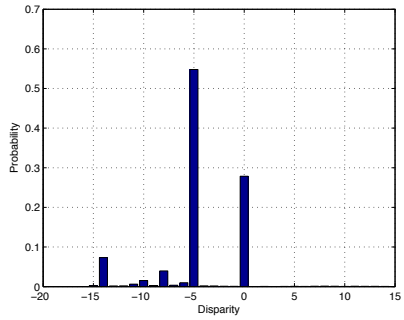


Figure 8: *Histogramme des disparités estimées par l'algorithme à M-chemins (PSNR=38.12 dB ; H=1.92 bpd).*

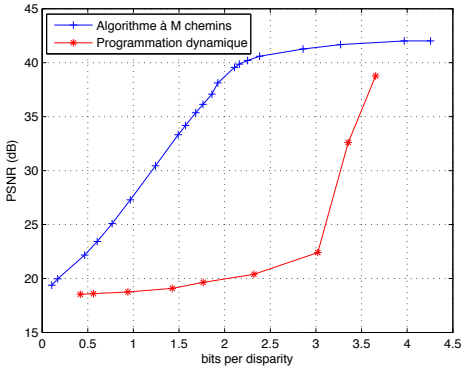


Figure 9: *Courbes débit-distorsion.*

Les courbes de la Figure 17 comparent les performances en termes de débit-distorsion. Les paramètres de notre algorithme ont été fixés comme suit : $M = 2$ et $\beta = 0.2$. En ce qui concerne la programmation dynamique, la taille des blocs varie de 1×1 à 17×17. Nous constatons, pour un débit équivalent de l'ordre de 2.54 bpd, qu'un gain de 12 dB est atteint en faveur de notre méthode d'optimisation.

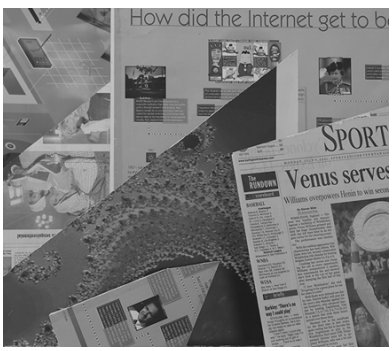


Figure 10: *Image originale de la vue droite.*

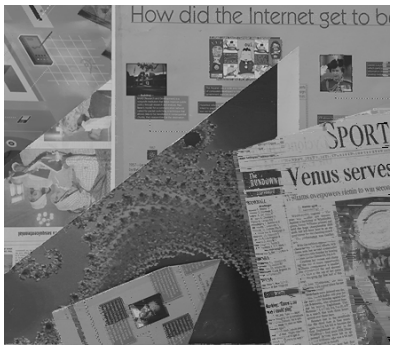


Figure 11: *Image reconstruite : carte de disparité estimée par l'algorithme de programmation dynamique (PSNR=27.06 dB ;H=4.46 bpd).*

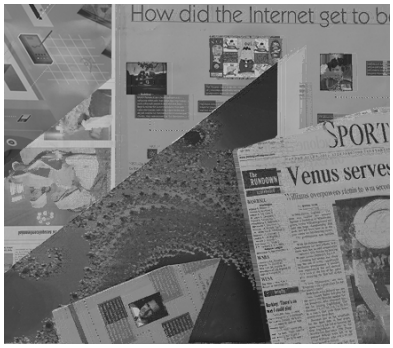


Figure 12: *Image reconstruite : carte de disparité estimée par l'algorithme à M-chemins (PSNR=27.39 dB ;H=1.85 bpd).*

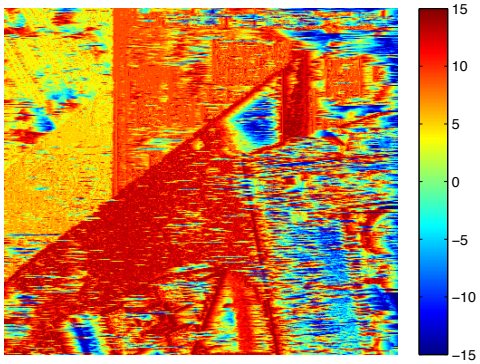


Figure 13: *Carte de disparité estimée par la programmation dynamique (PSNR=27.06 dB ; H=4.46 bpd).*

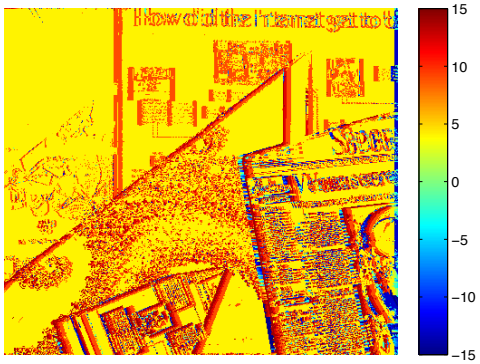


Figure 14: *Carte de disparité estimée par notre algorithme à M-chemins (PSNR=27.39 dB ;H=1.85 bpd).*

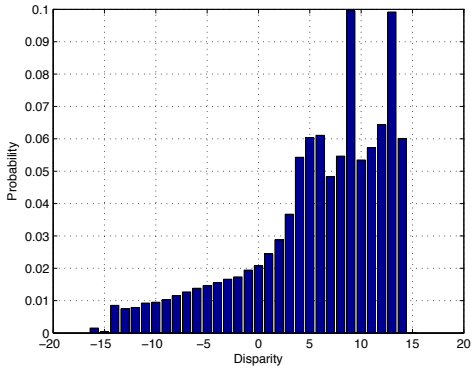


Figure 15: *Histogramme des disparités estimées par la programmation dynamique (PSNR=27.06 dB ; H=4.46 bpd).*

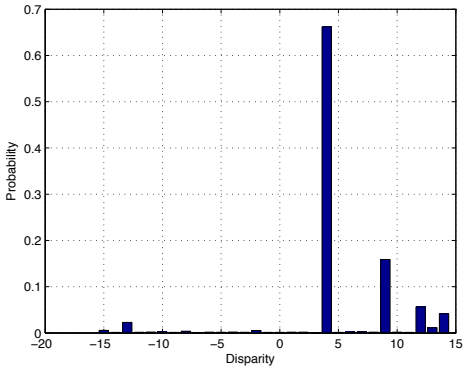


Figure 16: *Histogramme des disparités estimées par l'algorithme à M-chemins (PSNR=27.39 dB ; H=1.85 bpd).*

4. Conclusion

Nous avons développé un nouvel algorithme de mise en correspondance d'image stéréoscopique. Notre algorithme repose sur une métrique conjointe tenant compte à la fois de la qualité de l'image reconstruite (distorsion) mais également de l'entropie de la carte de disparité. Ce problème s'appuie sur le formalisme de Lagrange qui est résolu par la construction séquentielle d'un arbre à M-chemins. Cette stratégie permet non seulement de réduire la complexité de calcul mais également d'offrir de bons résultats en termes d'entropie-distorsion comparé aux résultats fournis par l'algorithme

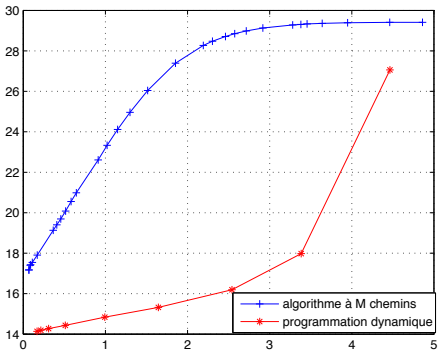


Figure 17: *Courbes débit-distorsion.*

classique de programmation dynamique. Des investigations sont en cours afin d'adapter notre algorithme à une mise en correspondance par bloc de façon à réduire significativement le débit tout en cherchant à préserver une bonne qualité de reconstruction.

Références

[BBH03] BROWN M. Z., BURSCHKA D., HAGER G. D. : Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 25*, Num. 8 (2003).

[BG05] BLEYER M., GELAUTZ M. : Graph-based surface reconstruction from stereo pairs using image segmentation. *Videometrics VIII. Vol. SPIE-5665* (janvier 2005).

[BVZ01] BOYKOV Y., VEKSLER O., ZABIH R. : Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI. Vol. 23*, Num. 11 (novembre 2001).

[Jel69] JELINEK F. : Fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development. Vol. 13*, Num. 6 (novembre 1969).

[Mid] :. *http ://vision.middlebury.edu/stereo/data/.*

[Nas92] NASRABADI N. M. : A stereo vision technique using curve-segments and relaxation matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI. Vol. 14*, Num. 5 (mai 1992).

[OK85] OHTA Y.-M., KANADE T. : Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 7*, Num. 2 (mars 1985).

[ref] :. *Matlab toolbox : http ://www.mathworks.fr/products/computer-vision/code-examples.html ?file=%2Fproducts%2Fdemos %2Fshipping%2Fvision%2Fvideostereo.html.*

[SS02] SCHARSTEIN D., SZELISKI R. : A taxonomy and evaluation of dense two-frame stereo correspondance algorithms. *International Journal of Computer Vision. Vol. 47*, Num. 1 (2002).

[Sun03] SUN J. : Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI. Vol. 25*, Num. 7 (décembre 2003).

[TCC07] TSENG Y.-C., CHANG N., CHANG T.-S. : Low memory cost block-based belief propagation for stereo correspondance. *IEEE International Conference on Multimedia and Expo* (2007).

[Vek05] VEKSLER O. : Stereo correspondence by dynamic programming on a tree. *IEEE Conference Proceedings of Computer Vision and Pattern Recognition. Vol. 2* (juin 2005).

Compression de contenu vidéo Super Multi-Vue avec parallaxe horizontale et verticale

A. Dricot[1,2], J. Jung[1], M. Cagnazzo [2], B. Pesquet [2], et F. Dufaux [2]

[1]Orange Labs
[2]Institut Mines-Télécom;
Télécom ParisTech; CNRS LTCI

Résumé

La vidéo Super Multi-Vue (SMV) est une technologie clé pour mettre en place les futurs services de vidéo 3D. Le SMV permet une visualisation sans lunette et élimine beaucoup des causes d'inconfort présentes dans les technologies de vidéo 3D actuelles. Le contenu vidéo SMV est composé de dizaines ou de centaines de vues d'une scène, qui peuvent être alignées soit uniquement dans la direction horizontale, soit dans les directions horizontale et verticale. Cet article compare plusieurs schémas de codage, puis propose une structure de codage qui exploite les corrélations inter-vues dans les deux directions, permettant de réduire le débit (pour une qualité donnée) de 29.1% *par rapport à une structure de référence basique. De plus, une amélioration des outils de codage Neighboring Block Disparity Vector (NBDV) et Inter-View Motion Prediction (IVMP) est proposée afin d'exploiter efficacement les structures de codage en deux dimensions, avec une réduction de débit allant jusqu'à* 4.2% *par rapport à l'encodeur référence 3D-HEVC.*

Super Multi-View video (SMV) is a key enabler for future 3D services. SMV allows a glasses-free viewing and eliminate many causes of discomfort existing in current 3D video technologies. SMV content consists of tens or hundreds of views of a scene, that can be aligned along the horizontal direction or both along horizontal and vertical directions. This paper compares several coding schemes, and proposes a coding structure that exploits inter-view correlations in horizontal and also vertical directions. This new structure provides a rate reduction (for the same quality) up to 29.1% *when compared to a basic anchor structure. Neighboring Block Disparity Vector (NBDV) and Inter-View Motion Prediction (IVMP) coding tools are further improved to efficiently exploit coding structures in two dimensions, with rate reduction up to* 4.2% *with respect to the reference 3D-HEVC encoder.*

Mots clé : Compression vidéo 3D, multi-vue, parallaxe de mouvement

1. Introduction

Le développement des technologies liées à la vidéo 3D tend à créer des expériences de visualisation de plus en plus immersives. Cependant, les technologies vidéo 3D actuellement disponibles sur le marché présentent plusieurs limitations [DPC13]. Avec la stéréoscopie 3D, le manque de confort dû au port des lunettes est combiné à des stimuli de perception qui ne sont pas naturels, comme le conflit entre vergence et accommodation, et qui peuvent causer des douleurs oculaires ainsi que des migraines. Avec les systèmes auto-stéréoscopiques sans lunette, le nombre réduit de vues ne permet pas d'avoir une parallaxe de mouvement fluide (c'est à dire que la visualisation n'est pas continue quand on bouge devant l'écran) et restreint la taille de la zone de visualisation, ce qui altère particulièrement la qualité et le confort de l'expérience de visualisation.

Une étude de la vidéo Super Multi-Vue (SMV) a été initiée pendant le meeting MPEG FTV d'Octobre 2013 [TSO*13a]. Le SMV utilise des dizaines ou des centaines de vues afin de créer une représentation du *light-field* d'une scène. Le light-field représente en principe tous les rayons de lumière dans une scène en 3D. C'est donc une fonction de deux angles (direction du rayon) et de trois coordonnées spatiales. Cette fonction en 5 dimensions est la fonction plénoptique [AB91]. Beaucoup des artefacts existants dans les technologies 3D actuelles peuvent être éliminés avec une représentation en light-field, en particulier le conflit vergence-accommodation. Elle permet une visualisation réaliste, sans lunette, et avec une parallaxe de mouvement (qui est un élément clé dans la perception du relief) fluide dans la direction horizontale et potentiellement dans la direction verticale. Plusieurs entreprises ont déjà démontré de l'intérêt pour le SMV en travaillant sur des écrans et des systèmes d'affichage dits light-field. La visio-conférence immersive est présentée comme un cas d'utilisation cible typique, ainsi que la diffusion live d'évènements sportifs en 3D, comme les Jeux Olympiques de 2020 au Japon, qui pourrait être filmés

par des ensembles de caméras et projetés sur des écran SMV géants dans les lieux publics de plusieurs grandes villes dans le monde [TSO*13c]. Il existe donc une demande et un besoin pour des nouvelles technologies de codage efficaces qui peuvent traiter la grande quantité de données nécessaire pour le SMV [TSO*13b].

Les extensions multi-vues des encodeurs standards peuvent permettre d'encoder du contenu SMV avec parallaxe horizontale. Des modifications de ces encodeurs sont proposées dans la littérature scientifique pour encoder du contenu avec parallaxe horizontale et verticale. Les méthodes de l'état de l'art présentent cependant des limitations dans leur utilisation des deux dimensions pour la prédiction inter-vues. On propose ici un schéma de prédiction inter-vues efficace pour exploiter les dimensions horizontale et verticale au niveau de la structure de codage. On propose ensuite des améliorations d'outils de codage inter-vues pour exploiter les structures en deux dimensions également au niveau pixellique.

Le reste de cet article est organisé comme suit. Des méthodes d'acquisition d'un ensemble de vues avec parallaxe horizontale et verticale sont décrites dans la Section 2. La Section 3 décrit des méthodes de l'état de l'art pour le codage de ce contenu SMV avec parallaxe dans les deux dimensions. Dans la Section 4, on décrit le schéma de prédiction inter-vues proposé et on montre des résultats expérimentaux contre les schémas de l'état de l'art. Les outils de codage inter-vues améliorés et adaptés à la parallaxe en deux dimensions sont proposés dans la section 5, qui contient également des résultats expérimentaux. La Section 6 conclut finalement cet article.

2. Acquisition et représentation du light-field

Une représentation du light-field d'une scène peut être obtenues à partir de plusieurs images captées depuis différents points/angles de vue. On considère ici deux technologies qui permettent d'obtenir du contenu avec parallaxe dans les directions horizontale et verticale : l'acquisition multi-caméras et l'imagerie intégrale.

Un contenu SMV peut être filmé avec une matrice de caméra comme le montre la Figure 1. Ces caméras peuvent être alignées horizontalement (donnant une parallaxe horizontale uniquement) ou horizontalement et verticalement dans le cas du contenu avec parallaxe en deux dimensions. Ce système de caméra peut être arrangé en linéaire, en arc ou même de manière aléatoire. Chaque caméra capte la scène d'un point de vue différent et le contenu résultant est un ensemble de vues avec des disparités horizontales et verticales.

L'imagerie intégrale (ou holoscopie) est une technique basée sur la photographie plénoptique [Lip08]. L'acquisition plénoptique est basée sur l'utilisation d'un panneau lenticulaire placé devant une caméra. Ce dispositif est schématisé dans la Figure 1. Le panneau lenticulaire est composé d'un grand nombre de microlentilles, pouvant avoir une forme ronde ou carrée, et pouvant être alignées en grille ou en quinconce. L'image holoscopique résultant de cette captation est une matrice de Micro-Images (MIs). Chaque microlentille produit une MI, et chaque MI contient de l'information sur la scène provenant de différents angles de vues. La connexion entre une image intégrale et un ensemble de vue d'une scène peut être tracée par l'extraction de vues, comme le montre

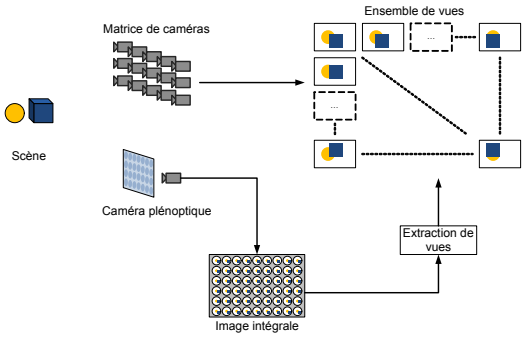


Figure 1: *Deux méthodes d'acquisition d'un light-field*

la Figure 1. Plusieurs méthodes sont proposées dans l'état de l'art pour extraire des vues à partir d'une image intégrale [GL10] [Lin13].

L'imagerie intégrale et l'acquisition SMV peuvent donner toutes les deux une représentation light-field car elles font un échantillonnage du light-field d'une scène en captant des images de cette scène selon plusieurs angles de vues. Cependant, cet échantillonnage est fait de deux façons distinctes, impliquant les compromis suivants. Un ensemble de caméras permet d'obtenir une *baseline* plus large (avec par exemple une distance de plusieurs mètres entre les deux extrémités du système de caméras) qu'une caméra holoscopique dont la baseline est limitée par les dimensions du panneau lenticulaire. Avec une caméra plénoptique, la résolution des vues extraites est limitée car un seul capteur est partagé par toutes les vues captées, alors qu'avec un ensemble de caméra, la résolution totale de chaque caméra est disponible pour chaque vue. De plus, les caméras plénoptiques permettent un échantillonnage plus dense du light-field, car la distance entre les caméras d'un ensemble de caméras est limitée par leurs tailles.

3. État de l'art

3.1. Extensions multi-vues des encodeurs vidéo standards

SMV définit du contenu vidéo 3D multi-vues avec des dizaines ou des centaines de vues, avec une parallaxe soit uniquement horizontale, soit dans les deux dimensions. Le nombre massif de vues augmente la quantité de données à traiter par rapport aux technologies vidéo 3D actuelles. La corrélations inter-vues est également augmentée. Les encodeurs multi-vues standards actuels sont conçus et implémentés pour un contenu avec parallaxe horizontale uniquement et avec un nombre limité de vues. MVC et MV-HEVC sont respectivement les extensions multi-vues des standards H.264/AVC et HEVC [Ohm13]. Ces extensions introduisent de la syntaxe haut-niveau permettant la prédiction inter-vues. L'extension 3D-HEVC fournit des outils de codage liés aux cartes de profondeur, des outils de prédictions inter-composantes (c'est à dire entre texture et carte de profondeur) et de nouveaux outils de codage pour les vues aux niveaux des *Coding Unit* (dans HEVC les CUs remplacent les *macroblocks* de H.264/AVC).

Dans la version du logiciel de référence utilisée pour nos

expériences (HTM7.0), la définition suivante s'applique. *Neighboring Block Disparity Vector* (NBDV) [ZCM12] et *Inter-View Motion Prediction* (IVMP) [TKCY12] sont des outils spécifiques à l'extension 3D-HEVC, conçus pour l'encodage multi-vues horizontal classique. Pour le CU courant, NBDV cherche un vecteur de disparité (DV) dans des CUs voisins (temporaux et spatiaux) déjà codés. Le DV dérivé par NBDV est utilisé par IVMP pour créer l'*Inter-view Predicted Motion Candidate* (IPMC). IPMC correspond aux paramètres de mouvement (vecteurs de mouvement et images de référence temporelle) du CU pointé par le DV dans la vue de référence. IPMC est inséré à la première place dans la liste de candidats du mode *Merge* [HOB*12]. Finalement, le DV lui-même est également inséré dans la liste du Merge en tant que candidat *Disparity Motion Vector* (DMV).

3.2. Améliorations pour les configurations avec parallaxe en deux dimensions

La première approche considérée pour encoder du contenu SMV avec parallaxe en deux dimensions est l'utilisation d'un encodeur multi-vue standard, avec une adaptation au niveau de la structure des références inter-vues. Dans [SGM11], les vues sont d'abord scannées en spirale, comme illustré dans la Fig. 2 (a), puis réalignées horizontalement. L'arrangement horizontal est ensuite encodé par un encodeur MVC en utilisant une structure de prédiction IBP (b). La Figure 2 (c) montre le schéma résultant de la représentation sur deux dimensions de cette structure IBP. Le principal inconvénient de cette approche est l'introduction de prédictions incohérentes et inefficaces.
Dans [MSMW07], il est proposé d'appliquer des structures horizontales en IPP ou IBP (Fig. 5(e) and (f)) à chaque ligne de la matrices de vues, et d'ajouter une prédiction inter-vue verticale uniquement sur la première colonne de vues ou sur la centrale, comme illustré dans les Fig. 3 (a),(b) and (c). Le nombre de prédictions inter-vues verticales utilisées est limité dans de telles structures.
Dans [CSK08], [CJSK09] et [CJSK10], une autre structure est proposée, illustrée dans la Fig. 3 (d). Chaque ligne de vues utilise une structure horizontale IBP et des prédictions inter-vues verticales sont ajoutées, donnant des vues de types : B1 qui ont uniquement deux références horizontales ou deux références verticales, B2 avec une référence horizontale et deux références verticales, et B3 avec deux références dans chaque direction. Les principales limites sont le nombre réduit de vues qui utilise une combinaison de références horizontales et verticales (moins de la moitié des vues sont de types B2 ou B3) et la distance entre la vue courante/à coder et sa vue de référence.
Une seconde approche au niveau CU est considérée dans [ADCL*12] et dans [CSK08], [CJSK09] et [CJSK10]. Des méthodes similaires sont proposées, basées sur la prédiction d'un DV pour la vue courante par interpolation des DVs des vues voisines.

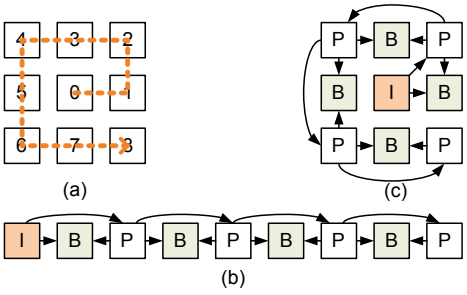


Figure 2: *Méthode de l'état de l'art [SGM11] pour 9 vues (a) scan en spirale, (b) structure de prédiction IBP, (c) schéma IBP équivalent en 2 dimensions*

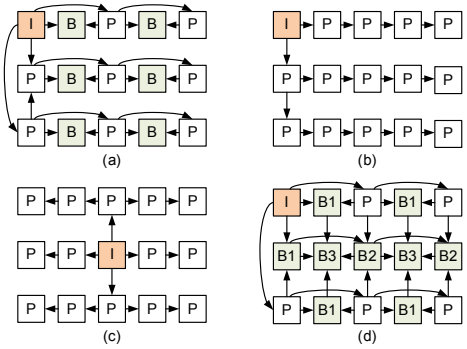


Figure 3: *Structures de l'état de l'art (a),(b) and (c) proposées dans [MSMW07], et (d) proposée dans [CSK08, CJSK09, CJSK10].*

4. Proposition d'une nouvelle structure de prédiction inter-vues

4.1. Schémas de référence et schéma proposé

On propose une structure pour les images de référence inter-vues, nommée *Central2D* et illustrée dans la Fig. 4 (b), qui permet d'exploiter efficacement un alignement de vues en deux dimensions. Pour une configuration avec $N \times M$ vues, le schéma *Central2D* est construit comme suit. La vue centrale est d'abord codée sans référence inter-vues. Les $N-1$ (respectivement $M-1$) vues qui sont sur le même axe horizontal (resp. vertical) que la vue centrale sont ensuite codées avec une référence inter-vue, étant la vue la plus proche dans la direction du centre. Toutes les autres vues sont codées en utilisant une référence inter-vues horizontale et une verticale, étant les vues les plus proches dans la direction du centre. Le schéma permet donc d'utiliser une combinaison de références horizontale et verticale pour un grand nombre de vues (seulement $M+N-1$ vues n'utilisent pas une référence dans chaque dimension). De plus, cette méthode minimise la distance entre la vue courante/à coder et ses images

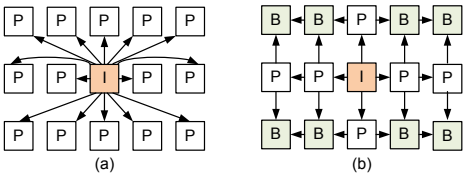


Figure 4: *(a) ancrage basique, (b) shéma proposé Central2D*

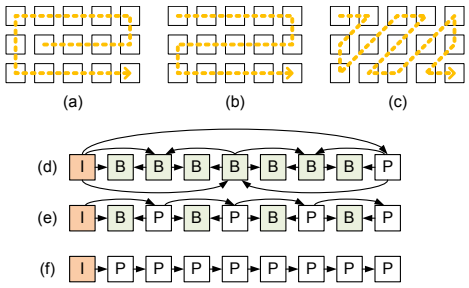


Figure 5: *Ordre de scan : (a) spirale, (b) perpendiculaire, (c) diagonale et structures horizontales d'images de référence inter-vue : (d) hiérarchique, (e) IBP, (f) IPP*

de référence inter-vues, et n'utilise pas de référence diagonale.
Dans la section suivante, le schéma proposé est comparé à une structure de référence basique (voir Fig. 4 (a)) avec seulement la vue centrale comme image de référence inter-vues pour toutes les autres vues. Cette comparaison permet de déterminer le bénéfice d'une prédiction inter-vues dans les deux directions et avec une distance réduite entre la vue à coder et sa vue de référence. Les structures tirées de l'état de l'art sont aussi testées dans nos expérimentations : [MSMW07] et [CSK08] correspondent aux schémas illustrés dans la Fig. 3 (c) and (d). [SGM11] correspond au scan en spirale avec une structure IBP (Fig. 2). Dans un but de comparaison, on propose également d'étendre la méthode [SGM11] en faisant varier l'ordre de scan et la structure comme c'est illustré dans la Fig. 5.

4.2. Résultats expérimentaux

Dans cette section, on teste les schémas de l'état de l'art et le schéma proposé avec MV-HEVC. La structure de prédiction temporelle reste telle que décrite dans les *Common Test Conditions* (CTC) [RMV12]. Les expérimentations sont effectuées avec le logiciel de référence de MV-HEVC dans sa version 7.0 (HTM7.0, avec la macro QC_MVHEVC activée). Deux séquences sont testées : *CoastalGuard* (50 frames, *computer generated*, de résolution 768×384) et *Akko&Kayo* (290 frames, filmée, de résolution 640×480). Des configurations de 3×3 vues et de 11×5 vues sont testées. Les résultats sont mesurés en utilisant la métrique dite Bjøntegaard Delta rate (BD-rate) [Bjø01], sur les QPs (*Quantization Parameter*) 22-27-32-37. Cette métrique mesure la variation de débit à qualité égale, par rapport à une technique de référence. Dans nos expériences, la référence est le schéma d'ancrage basique (Fig. 4 (a)). Les valeurs négatives représentent un gain, c'est-à-dire, une réduction de débit par rapport à la référence.
Le Tableau 1 montre que pour les deux séquences, avec la configuration 3×3 vues, le schéma *Central2D*, la méthode [MSMW07] et la structure IPP associée aux scans perpendiculaire et spirale sont plus efficaces que les autres méthodes. Ces schémas n'utilisent pas de référence inter-vues en diagonale et minimisent la distance entre la vue à coder et la référence inter-vues. Le gain supplémentaire pour *Central2D* est dû à l'utilisation de références inter-vues dans les deux directions (horizontale et verticale). Le Tableau 2 montre que le schéma *Central2D* reste la structure la plus cohérente et efficace avec un plus grand nombre de vues.
Le gain final en BD-rate apporté par la structure *Central2D* sur l'ancrage basique monte jusque 8.2% and 29.1% dans les configurations de 3×3 et 11×5 vues respectivement.

Coast 3×3			
	spirale	perpendiculaire	diagonale
IPP	-1.2%	-2.2%	5.1%
IBP	9.1%	7.1%	11.4%
Hiérarchique	3.0%	4.4%	8.4%
Méthode [CSK08]	2.1%		
Méthode [MSMW07]	-6.8%		
CENTRAL2D	-7.1%		
Akko 3×3			
	spirale	perpendiculaire	diagonale
IPP	-4.9%	-5.5%	8.8%
IBP	2.7%	-4.0%	-1.9%
Hiérarchique	1.9%	2.4%	4.0%
Méthode [CSK08]	7.8%		
Méthode [MSMW07]	-7.7%		
CENTRAL2D	-8.2%		

Table 1: *Variations du BD-rate pour les structures de l'état de l'art et proposée comparées à l'ancrage basique - avec* 3×3 *vues*

Coast 11×5			
	spirale	perpendiculaire	diagonale
IPP	-20.5%	-19.6%	16.1%
IBP	-15.9%	-14.9%	-13.9%
Hiérarchique	-8.4%	-9.3%	-13.0%
Méthode [CSK08]	-19.5%		
Méthode [MSMW07]	-24.4%		
CENTRAL2D	-29.1%		
Akko 11×5			
	spirale	perpendiculaire	diagonale
IPP	-22.9%	-24.8%	-6.5%
IBP	-20.0%	-23.4%	-2.4%
Hiérarchique	-14.9%	-20.2%	-3.7%
Méthode [CSK08]	-24.2%		
Méthode [MSMW07]	-25.9%		
CENTRAL2D	-27.6%		

Table 2: *Variations du BD-rate pour les structures de l'état de l'art et proposée comparées à l'ancrage basique - avec* 11×5 *vues*

5. Adaptation et amélioration d'outils de codage inter-vues

5.1. Amélioration de la liste de candidats du mode Merge

On propose dans cette section une modification normative des outils de codage NBDV et IVMP. NBDV et IVMP sont des outils de codage spécifiques implémentés de manière à fonctionner dans les conditions de tests standards (Common Test Conditions - CTC [RMV12]), c'est à dire avec une seule image de référence inter-vue horizontale, étant la vue centrale (*baseview*, avec l'indice 0). On adapte ici ces outils

en permettant l'utilisation de plusieurs images de références inter-vues, avec des indices différents de 0, et pouvant être horizontales ou verticales.
En plus de cette adaptation, l'amélioration suivante est proposée. Lorsqu'une vue B utilisant une référence inter-vue horizontale et une verticale est encodée, la version modifiée de NBDV cherche deux DVs (un pour chaque image de référence inter-vues). La recherche du second DV seule n'apporte aucun gain en elle-même, mais va être utilisée pour les candidats IPMC et DMV. Le nouveau second DV est utilisé pour insérer un second IPMC à la seconde place de la liste des candidats du mode Merge. Pour le candidat DMV, le couple de DVs est utilisé, permettant de faire une prédiction bidirectionnelle (ou bi-prédiction).

5.2. Dérivation inter-vues du second DV

On propose d'augmenter les chances de trouver un second DV avec NBDV, afin d'augmenter l'efficacité des candidats IPMC et DMV modifiés. Les étapes sont illustrées dans la Fig. 6. Pour la vue courante, NBDV doit d'abord trouver un premier DV horizontal, pointant sur un CU de référence dans la vue de référence. Si cette vue de référence horizontale a une vue de référence verticale, et si le CU de référence est codé par prédiction inter-vues, le DV vertical utilisé pour la prédiction est hérité/dérivé (par simple copie) en tant que second DV pour le CU courant, et est ensuite utilisé pour les candidats IPMC et DMV (de la manière décrite dans la section précédente). On note que cette méthode peut être utilisée pour les vues de type B ayant une référence inter-vues horizontale et une verticale, ce qui rend la structure *Central2D* précédemment proposée la plus adéquate pour ces outils de codage modifiés.

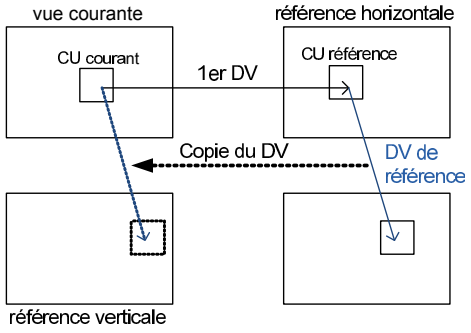


Figure 6: *Dérivation inter-vues du second DV*

5.3. Résultats expérimentaux

Dans cette section, on teste les modifications proposées des outils de codage NBDV et IVMP. Les expérimentations sont effectuées avec le logiciel de référence de 3D-HEVC dans sa version 7.0 (HTM7.0). Les conditions de test sont les mêmes que dans la Section 4.2 (permettant l'utilisation de structures de codage en deux dimensions). La structure *Central2D* précédemment proposée est utilisée pour tous les tests dans la suite. La référence est 3D-HEVC (HTM7.0 sans modification algorithmique).

Référence : 3D-HEVC (HTM7.0 sans modification algorithmique)				
	3×3 vues		11×5 vues	
	Coast	Akko	Coast	Akko
Adaptation seule	-1.1%	-2.3%	-2.4%	-3.3%
BiDMV	-1.2%	-2.4%	-2.7%	-3.7%
2 IPMC	-1.1%	-2.3%	-2.8%	-3.5%
Combinaison	-1.3%	-2.5%	-3.1%	-3.9%

Table 3: *Variations du BD-rate pour les outils NBDV et IVMP améliorés, utilisant un DV pour chaque image de référence inter-vues*

Référence : 3D-HEVC (HTM7.0 sans modification algorithmique)				
	3×3 vues		11×5 vues	
	Coast	Akko	Coast	Akko
BiDMV + dérivation	-1.9%	-2.9%	-3.4%	-3.9%
2 IPMC + dérivation	-1.3%	-2.4%	-2.8%	-3.5%
Combinaison + dérivation	-2.0%	-2.9%	-3.9%	-4.2%

Table 4: *Variations du BD-rate pour les outils NBDV et IVMP améliorés, avec la dérivation inter-vues du second DV*

Le Tableau 3 montre que l'adaptation de NBDV et IVMP aux structures bidimensionnelles apporte un gain (BD-rate) jusque 3.3%, ce qui confirme l'impact de l'utilisation des deux dimensions horizontale et verticale au niveau de la structure des images de référence inter-vues. L'insertion d'un second IPMC dans la liste des candidats du mode Merge et la bi-prédiction pour le candidat DMV apporte séparément des gains jusque 2.4% pour la configuration 3×3 vues et 3.7% pour la configuration 11×5 vues. La combinaison des deux améliorations apporte des gains jusque 2.5% and 3.9% respectivement avec 3×3 et avec 11×5 vues. Le résultat pour la combinaison des deux outils est légèrement supérieur à la somme des deux pris séparément car la bi-prédiction permet à NBDV de trouver plus souvent un second DV, et donc augmente les chances d'avoir un second IPMC efficace.
Le Tableau 4 montre que la dérivation inter-vues proposées pour le second DV est efficace et améliore la performance du codage de la méthode proposée complète (incluant l'adaptation de NBDV et IVMP aux structures bidimensionnelles, les deux IPMC, la bi-prédiction pour DMV et la dérivation inter-vue du second DV) jusque 2.9% et 4.2% pour la séquence *Akko&Kayo* respectivement avec 3×3 et avec 11×5 vues.

6. Conclusion

On propose dans cet article une structure pour les images de référence inter-vues adaptée au contenu vidéo 3D *light-field* avec une parallaxe de mouvement en horizontale et en verticale (c'est à dire avec des vues alignées dans les deux directions). La principale caractéristique de cette structure est la distance réduite entre la vue à coder et sa vue de référence, ainsi que l'utilisation de références inter-vues horizontales et verticales. Le schéma proposé surpasse l'ancrage basique avec un gain allant jusque 29.1% (en BD-rate), montrant l'impact de l'utilisation efficace des deux directions horizontale et verticale dans le schéma des images de référence inter-vues. On propose également d'améliorer les outils de

codage NBDV et IVMP (dans 3D-HEVC) afin d'exploiter les directions horizontale et verticale, avec un gain allant jusque 4.2%. Les résultats des méthodes proposées montre qu'exploiter efficacement les deux dimensions horizontale et verticale d'un contenu SMV avec parallaxe dans les deux dimensions au niveau de la structure de codage et au niveau des outils de codage permet d'améliorer de manière significative la performance de compression.

Références

[AB91] ADELSON E. H., BERGEN J. R. : The plenoptic function and the elements of early vision. *Computational models of visual processing. Vol. 1*, Num. 2 (1991).

[ADCL*12] AVCI A., DE COCK J., LAMBERT P., BEERNAERT R., DE SMET J., BOGAERT L., MEURET Y., THIENPONT H., DE SMET H. : Efficient disparity vector prediction schemes with modified P frame for 2D camera arrays. *Journal of Visual Communication and Image Representation. Vol. 23*, Num. 2 (February 2012), 287–292.

[Bjø01] BJØNTEGAARD G. : Calculation of average PSNR differences between RD-curves. In *VCEG Meeting* (Austin, USA, April 2001).

[CJSK09] CHUNG T.-Y., JUNG I.-L., SONG K., KIM C.-S. : Virtual view interpolation and prediction structure for full parallax multi-view video. In *Advances in Multimedia Information Processing - PCM*, vol. 5879. Springer, 2009, pp. 543–550.

[CJSK10] CHUNG T.-Y., JUNG I.-L., SONG K., KIM C.-S. : Multi-view video coding with view interpolation prediction for 2D camera arrays. *Journal of Visual Communication and Image Representation. Vol. 21*, Num. 5 (July-August 2010), 474–486.

[CSK08] CHUNG T., SONG K., KIM C.-S. : Compression of 2-D wide multi-view video sequences using view interpolation. In *15th IEEE International Conference on Image Processing (ICIP)* (San Diego, CA, USA, October 2008), IEEE, pp. 2440–2443.

[DPC13] DUFAUX F., PESQUET-POPESCU B., CAGNAZZO M. : *Emerging technologies for 3D video : content creation, coding, transmission and rendering*. Wiley Eds, 2013.

[GL10] GEORGIEV T., LUMSDAINE A. : Focused plenoptic camera and rendering. *Journal of Electronic Imaging. Vol. 19*, Num. 2 (2010), 021106.

[HOB*12] HELLE P., OUDIN S., BROSS B., MARPE D., BICI M. O., UGUR K., JUNG J., CLARE G., WIEGAND T. : Block merging for quadtree-based partitioning in hevc. *IEEE Transactions on Circuits and Systems for Video Technology. Vol. 22*, Num. 12 (December 2012), 1720–1731.

[Lin13] LINO J. F. O. : 2D image rendering for 3D holoscopic content using disparity-assisted patch blending. *Thesis to obtain the Master of Science Degree* (October 2013).

[Lip08] LIPPMANN G. : Epreuves reversibles donnant la sensation du relief. *J. Phys. Theor. Appl.. Vol. 7*, Num. 1 (1908), 821–825.

[MSMW07] MERKLE P., SMOLIC A., MULLER K., WIEGAND T. : Efficient prediction structures for multiview video coding. *IEEE Transactions on Circuits and Systems for Video Technology. Vol. 17*, Num. 11 (November 2007), 1461–1473.

[Ohm13] OHM J.-R. : Overview of 3D video coding standardization. In *International Conference on 3D Systems and Applications* (Osaka, Japan, June 2013).

[RMV12] RUSANOVSKY D., MULLER K., VETRO A. : Common test conditions of 3DV core experiments. In *International Organisation For Standardisation* (October 2012), ITU-T SG 16 WP 3 & ISO/IEC JTC1/SC29/WG11 JCT3V-B11000.

[SGM11] SHI S., GIOIA P., MADEC G. : Efficient compression method for integral images using multi-view video coding. In *18th IEEE International Conference on Image Processing (ICIP)* (Brussels, Belgium, September 2011), IEEE, pp. 137–140.

[TKCY12] TECH G., K.WEGNER, CHEN Y., YEA S. : 3D-HEVC test model 2. In *International Organisation For Standardisation* (October 2012), ITU-T SG 16 WP 3 & ISO/IEC JTC1/SC29/WG11 JCT3V-B1005.

[TSO*13a] TEHRANI M. P., SENOH T., OKUI M., YAMAMOTO K., INOUE N., FUJII T. : [m31095][FTV AHG] Use cases and application scenarios for super multiview video and free-navigation. In *International Organisation For Standardisation* (October 2013), ISO/IEC JTC1/SC29/WG11.

[TSO*13b] TEHRANI M. P., SENOH T., OKUI M., YAMAMOTO K., INOUE N., FUJII T. : [m31103][FTV AHG] Introduction of super multiview video systems for requirement discussion. In *International Organisation For Standardisation* (October 2013), ISO/IEC JTC1/SC29/WG11.

[TSO*13c] TEHRANI M. P., SENOH T., OKUI M., YAMAMOTO K., INOUE N., FUJII T. : [m31261][FTV AHG] Multiple aspects. In *International Organisation For Standardisation* (October 2013), ISO/IEC JTC1/SC29/WG11.

[ZCM12] ZHANG L., CHEN Y., M.KARCZEWICZ : 3D-CE5.h related : Disparity vector derivation for multiview video and 3DV. In *International Organisation For Standardisation* (July 2012), ISO/IEC JTC1/SC29/WG11 MPEG2012/m24937.

Transformation d'un dispositif multimédia webcam-écran en un scanner 3D

Y. Quéau R. Modrzejewski P. Gurdjos J.-D. Durou

IRIT, UMR CNRS 5505, Toulouse

Résumé

Nous étudions un dispositif de scannage 3D constitué d'un couple webcam-écran, où l'écran est utilisé comme source lumineuse. Ceci permet de transformer en scanner 3D n'importe quel dispositif multimédia comprenant ces deux éléments (ordinateurs portables, smartphones, tablettes etc.). Un protocole d'étalonnage simplifié est introduit, pour lequel nous démontrons que deux prises de vue sont suffisantes. Une fois cet étalonnage géométrique effectué, nous montrons que le dispositif étudié permet d'effectuer la reconstruction 3D sans ambiguïté, grâce à la technique de stéréophotométrie.

We study a multimedia system composed of both a webcam and a screen, where the screen is considered as a light source. Such a system allows one to turn a laptop, a smartphone or a tablet into a 3D-scanner. We propose a simplified calibration procedure which requires as few as two input images, and demonstrate that such a system allows one to get a 3D-reconstruction without ambiguity, using the photometric stereo technique.

Mots clé : étalonnage, webcam, écran LCD, reconstruction 3D, stéréophotométrie.

1. Introduction

L'utilisation de dispositifs multimédia comprenant à la fois un écran et une webcam est devenue incontournable. Dans de nombreux terminaux numériques tels que les ordinateurs portables, les smartphones ou les tablettes, ces deux entités coexistent. Et lorsque cela n'est pas le cas, il est toujours possible d'attacher la webcam à l'écran par une pince (cf. figure 1).

En détournant l'écran de son usage standard pour le transformer en source lumineuse [FY07], on dispose d'un système de reconstruction 3D par stéréophotométrie : m images de la scène sont acquises sous le même point de vue (le système est supposé immobile, tout comme la scène à reconstruire), mais sous m éclairages différents, contrôlés par l'affichage de motifs connus sur l'écran du dispositif. Pour pouvoir caractériser géométriquement l'éclairage incident relativement à la scène, il est toutefois nécessaire d'avoir étalonné géométriquement le système, c'est-à-dire de connaître la pose de la webcam relativement à l'écran.

Notre contribution est double : nous introduisons d'abord une procédure d'étalonnage géométrique simplifiée, dans le cas particulier où le mouvement de la webcam est limité à une rotation autour d'un axe parallèle à l'axe horizontal de l'écran (cf. figure 1) ; nous montrons ensuite que, grâce à cet étalonnage, le dispositif multimédia étudié permet de garantir l'unicité de la solution au problème de la reconstruction 3D par stéréophotométrie, là où la grande majorité des méthodes existantes ne fournissent de solution qu'à une ambigüité près.

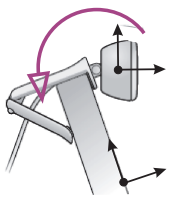


Figure 1: *Le dispositif consiste en un écran auquel est attachée une caméra (webcam). La caméra possède quatre degrés de liberté relativement à l'écran, qui correspondent aux trois coordonnées de son centre et à une rotation autour de l'axe horizontal de l'écran. (source : www.logitech.com/en-us/support/hd-webcam-c270).*

2. Étalonnage géométrique du système

Dans [RBN10], il est établi que l'image d'un objet obtenue par réflexion sur un miroir plan est géométriquement équivalente à l'image de cet objet obtenue par une caméra virtuelle située derrière le miroir, qui aurait les mêmes paramètres internes que la caméra réelle et dont la pose serait symétrique, par rapport au miroir, à celle de la caméra réelle

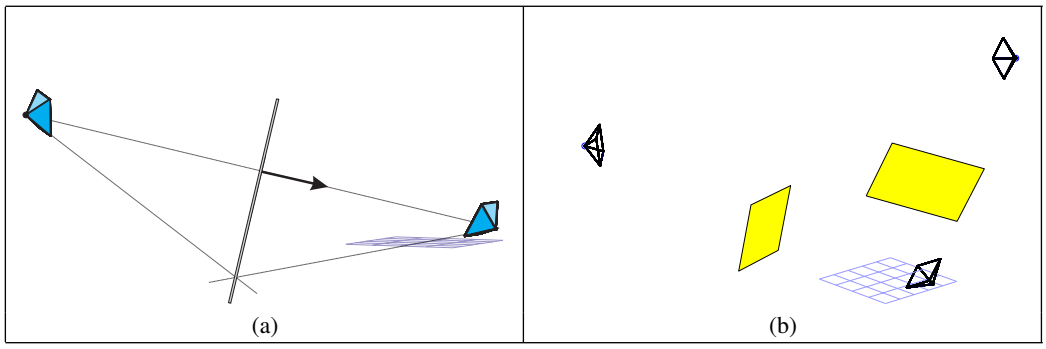


Figure 2: *(a) En plaçant un miroir face à la caméra réelle (à droite), on définit une caméra virtuelle (à gauche), symétrique de la caméra réelle par rapport au plan du miroir. (b) En utilisant plusieurs poses du miroir (en jaune), on définit autant de caméra virtuelles que de poses.*

(cf. figure 2-a). Nous allons utiliser cette propriété pour l'étalonnage géométrique du dispositif, et montrer qu'en utilisant le miroir pour réfléchir un motif affiché à l'écran, l'étalonnage géométrique du système (paramètres internes de la caméra et pose relativement à l'écran) peut être effectué analytiquement, à partir d'un minimum de deux prises de vue.

2.1. Notations

Dans ce travail, nous considérons que le repère tridimensionnel de référence est un repère orthonormé attaché à l'écran, orienté comme l'indique la figure 1 : l'axe des y est parallèle aux colonnes de l'écran, tandis que l'axe des z est orthogonal à l'écran.

Nous modélisons la caméra par un « trou d'épingle » de type CCD [HZ03]p.156, caractérisé par la position $W = (X_{\mathbf{w}}, Y_{\mathbf{w}}, Z_{\mathbf{w}})$ de son centre optique et l'orientation de l'axe optique. Motivés par l'observation des terminaux tels que les tablettes tactiles ou les webcams que l'on peut fixer sur un écran, nous nous limitons au cas suivant :

Hypothèse 1 *La caméra est fixée à l'écran par une pince, qui limite son mouvement à une rotation autour d'un axe parallèle à l'axe horizontal de l'écran (cf. figure 1).*

La caméra a donc quatre degrés de liberté qui correspondent à l'angle θ de la rotation autour de cet axe et aux trois coordonnées cartésiennes $X_{\mathbf{w}}, Y_{\mathbf{w}}, Z_{\mathbf{w}}$ de son centre optique.

D'après le modèle de caméra CCD, la webcam est décrite par sa matrice de projection, de la forme :

$$\mathsf{P} = \mathsf{K}\mathsf{R}\left[\mathsf{I} \mid -\mathbf{t}_{\mathbf{w}}\right] \tag{1}$$

où K est la matrice des paramètres internes de la webcam, I est la matrice unité et :

$$\mathsf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \quad \text{et} \quad \mathbf{t}_{\mathbf{w}} = \begin{bmatrix} X_{\mathbf{w}} \\ Y_{\mathbf{w}} \\ Z_{\mathbf{w}} \end{bmatrix} \tag{2}$$

Le miroir utilisé est supposé parfaitement plan. Ses n poses définissent n miroirs virtuels $\mathcal{M}_i$, représentés géométriquement par leurs plans de support. Le plan de support du miroir virtuel $\mathcal{M}_i$ est situé à la distance d_i de l'origine du repère considéré, et son vecteur normal est noté $\mathbf{n}_i$ ($\|\mathbf{n}_i\| = 1$). Il est donc caractérisé par le vecteur de coordonnées $[\mathbf{n}_i^\top, d_i]^\top$.

Enfin, nous définissons une caméra virtuelle par réflexion de la webcam par rapport à chaque pose du miroir, qui est décrite par une matrice de projection de la forme :

$$\mathsf{P}_i = \mathsf{K}_i\mathsf{R}_i\left[\mathsf{I} \mid -\mathbf{t}_{\mathbf{w}}^i\right] \tag{3}$$

Le système étudié a les propriétés suivantes, $\forall i \in [1,n]$:

$$\mathsf{K} = \mathsf{K}_i \tag{4}$$

$$\mathsf{P} = \mathsf{P}_i\mathsf{S}_i \tag{5}$$

où :

$$\mathsf{S}_i = \begin{bmatrix} \mathsf{I} - 2\mathbf{n}_i\mathbf{n}_i^\top & -2d\mathbf{n}_i \\ \mathbf{0}_3^\top & 1 \end{bmatrix} \tag{6}$$

est la matrice de la réflexion par rapport au miroir $\mathcal{M}_i$.

Le problème d'étalonnage que nous devons résoudre peut alors être formalisé de la façon suivante :

Problème 2 *À partir des poses des caméras virtuelles, sous l'hypothèse 1, déterminer la pose de la webcam dans un repère orthonormé attaché à l'écran.*

Nous allons montrer que le problème 2 admet une solution analytique unique, *à partir d'un minimum de deux poses* du miroir. En particulier, nous allons établir que la composante rotationnelle de la pose de la webcam est solution d'un système linéaire homogène, et que la composante translationnelle peut être obtenue en résolvant un problème d'intersection de droites.

2.2. Estimation des paramètres intrinsèques

Commençons par l'estimation de la matrice K des paramètres intrinsèques. Ces paramètres peuvent être estimés à partir de $m \geq 2$ images d'une mire d'étalonnage (par exemple un damier) posée sur un support plan, prises sous différentes orientations.

Le miroir peut être utilisé pour simuler ce protocole : une mire d'étalonnage affichée à l'écran est réfléchie sur le

miroir. En changeant l'orientation du miroir, tout se passe comme si l'on utilisait des mires virtuelles directement visibles depuis la caméra. Les méthodes classiques d'étalonnage peuvent alors être utilisées. Dans notre implémentation, nous utilisons la méthode de Bouguet [Bou04].

Les caméras virtuelles ayant les mêmes paramètres internes que la webcam, cf. (4), ceci permet également d'estimer les matrices K_i, ainsi que les poses des caméras virtuelles relativement au miroir (matrices R_i et vecteurs $\mathbf{t}_i$).

2.3. Estimation des paramètres extrinsèques

Intéressons-nous maintenant au calcul de la pose de la webcam, représentée par le produit $\mathsf{R}\left[\mathsf{I} \mid -\mathbf{t_w}\right]$.

Composante rotationnelle de la pose de la webcam. Il s'agit d'estimer la matrice R. Par souci de clarté, nous omettons momentanément les indices i pour les caméras virtuelles, car il est possible d'obtenir une solution unique en R pour chacune d'elles : l'estimation peut ensuite être rendue robuste en moyennant les résultats obtenus pour les différentes caméras virtuelles.

Considérons la sous-matrice obtenue en supprimant les troisièmes ligne et colonne de S :

$$\bar{\mathsf{S}} = \mathsf{I} - 2\mathbf{n}\mathbf{n}^\top \tag{7}$$

On montre facilement que :

$$\det \bar{\mathsf{S}} = -1 \tag{8}$$

De l'égalité (5), on déduit que :

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \sim \mathsf{R}\bar{\mathsf{S}} \tag{9}$$

où $\sim$ désigne l'égalité projective.

En exploitant l'égalité (9), on utilise pour estimer $\bar{\mathsf{S}}$ les six contraintes linéaires suivantes par rapport aux six élements de la matrice (symétrique) $\bar{\mathsf{S}}$:

$$\underbrace{\begin{bmatrix} 0 & -R_{21} & R_{31} & -R_{22} & R_{32}-R_{23} & R_{33} \\ 0 & R_{31} & R_{21} & R_{32} & R_{22}+R_{33} & R_{23} \\ 0 & R_{11} & 0 & R_{12} & R_{13} & 0 \\ 0 & 0 & R_{11} & 0 & R_{12} & R_{13} \\ R_{21} & R_{21} & R_{23} & 0 & 0 & 0 \\ R_{31} & R_{32} & R_{33} & 0 & 0 & 0 \end{bmatrix}}_{\mathsf{C}} \underbrace{\begin{pmatrix} S_{11} \\ S_{12} \\ S_{13} \\ S_{22} \\ S_{23} \\ S_{33} \end{pmatrix}}_{\mathbf{X}} = \mathbf{0}_6 \tag{10}$$

D'un point de vue théorique, le système obtenu est compatible et admet une solution exacte car la matrice C est, en général, une matrice d'ordre 6 et de rang 5. En présence de bruit sur les données, on a $\mathit{rang}(\mathsf{C}) > 5$, donc une solution au sens des moindres carrés doit être recherchée. Pour éviter la solution triviale $\mathbf{X} = \mathbf{0}_6$, on impose la contrainte $\|\mathbf{X}\| = 1$. Le problème s'écrit alors $\arg\min_{\mathbf{X}} \|\mathsf{C}\mathbf{X}\|^2$ s.c. $\|\mathbf{X}\| = 1$, qui peut être résolu par décomposition en valeurs singulières. Soit $\hat{\mathsf{S}}$ la solution pour $\bar{\mathsf{S}}$, normalisée de telle façon que (8) soit satisfaite. On obtient :

$$\cos\theta = (\mathsf{R}\hat{\mathsf{S}})_{22} \quad ; \quad \sin\theta = (\mathsf{R}\hat{\mathsf{S}})_{32} \tag{11}$$

en notant que $\mathsf{R}\hat{\mathsf{S}}$ est nécessairement de la forme du membre gauche de (9).

Composante translationnelle de la pose de la webcam. Il s'agit d'estimer la translation $\mathbf{t_w}$ du centre optique de la webcam. Ce dernier est situé à l'intersection des droites passant par les centres optiques des n caméras virtuelles, représentés par les vecteurs $\mathbf{t}_i$, et dirigées par les vecteurs normaux aux plans des miroirs correspondants, représentés par $\mathbf{n}_i$.

On connaît les vecteurs $\mathbf{t}_i$ et, en utilisant l'égalité (7), on estime les vecteurs $\mathbf{n}_i$ à partir de la décomposition en valeurs singulières de $\frac{1}{2}(\mathsf{I} - \hat{\mathsf{S}}_i)$:

$$\hat{\mathbf{n}}_i = \mathsf{U}(1,0,0)^\top \quad \text{si} \quad \mathsf{U}\mathbf{\Sigma}\mathsf{V}^\top = \frac{1}{2}(\mathsf{I} - \hat{\mathsf{S}}_i) \tag{12}$$

où U et V sont deux matrices orthogonales d'ordre 3 et $\mathbf{\Sigma}$ est la matrice diagonale des valeurs singulières. Il subsiste une ambiguïté sur le signe de $\hat{\mathbf{n}}_i$, qui peut être levée aisément puisqu'on sait que le miroir est orienté face à la caméra.

Chaque droite passant par le centre, de vecteur $\mathbf{t}_i$, d'une caméra virtuelle numéro i, et dirigée par le vecteur normal $\mathbf{n}_i$, peut être représentée par une matrice de Plücker [HZ03]p.70. En introduisant les vecteurs de coordonnées homogènes $\tilde{\mathbf{n}}_i = (\mathbf{n}_i^\top, 0)^\top$ et $\tilde{\mathbf{t}}_i = (\mathbf{t}_i^\top, 1)^\top$, cette matrice peut s'écrire :

$$\mathsf{L}_i = \tilde{\mathbf{t}}_i \tilde{\mathbf{n}}_i^\top - \tilde{\mathbf{n}}_i \tilde{\mathbf{t}}_i^\top \tag{13}$$

L_i est une matrice d'ordre 4, définie à un facteur près, antisymétrique avec une diagonale nulle, et de rang 2. Elle ne comporte donc que quatre degrés de liberté, comme toute droite de l'espace affine tridimensionnel.

On utilisera comme équations de base du problème d'intersection les équations linéaires fournies par le système homogène suivant :

$$\underbrace{\begin{bmatrix} \mathsf{L}_1^* \\ \vdots \\ \mathsf{L}_n^* \end{bmatrix}}_{\mathsf{D}} \underbrace{\begin{pmatrix} \mathbf{t_w} \\ 1 \end{pmatrix}}_{\tilde{\mathbf{Y}}} = \mathbf{0}_{4n\times 4} \tag{14}$$

où L_i^* désigne la seconde matrice de Plücker duale à L_i. Il s'ensuit qu'un minimum de deux droites est nécessaire pour estimer la position du centre de la webcam. Pour éviter la solution triviale $\tilde{\mathbf{Y}} = \mathbf{0}_4$, on impose la contrainte $\left\|\tilde{\mathbf{Y}}\right\| = 1$. Le problème s'écrit alors $\arg\min_{\tilde{\mathbf{Y}}} \left\|\mathsf{D}\tilde{\mathbf{Y}}\right\|^2$ s.c. $\left\|\tilde{\mathbf{Y}}\right\| = 1$, qui peut être résolu par décomposition en valeurs singulières.

2.4. Considérations expérimentales

À partir de n images du miroir, on peut retrouver la pose de la caméra réelle par rapport à l'écran. Nous résumons les différentes étapes de l'étalonnage géométrique dans l'algorithme suivant :

Données : n images d'un miroir réfléchissant un écran sur lequel est affichée une mire d'étalonnage.

Résultats :

— $(\mathsf{R}, \mathbf{t_w})$: pose de la caméra réelle.

— $[\mathbf{n}_i^T, d_i]^\top$, $i \in [1,n]$: vecteurs des n plans de support des miroirs virtuels.

1. Initialisation : estimer les paramètres internes de la caméra réelle et les poses $(\mathsf{R}_i, \mathbf{t}_i)$ des n caméras virtuelles [Bou04].

2. Composante rotationnelle :
   Pour $k = 1...n$, estimer θ_i depuis R_i, à partir du système linéaire (10) et de l'identification (11).
3. $\theta = \text{médiane}\{\theta_i\}$
4. $\mathsf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$
5. Composante translationnelle :
   Pour $k = 1...n$, calculer la matrice de Plücker L_k par l'équation (13), et sa seconde matrice duale L_k^*.
6. Résoudre le système (14) au sens des moindres carrés et en déduire $\mathbf{t_w}$.

En théorie, deux poses du miroir suffisent à étalonner géométriquement le système. En effet, il s'agit du nombre minimal d'images permettant de réaliser l'étalonnage des paramètres intrinsèques et des poses des caméras virtuelles [Bou04]. Ensuite, l'estimation de la composante rotationnelle peut être théoriquement menée à partir d'une seule pose, mais en pratique il est nécessaire d'en utiliser davantage, afin de garantir une certaine robustesse aux effets de bruit ou de flou de bougé : pour cela, nous effectuons une estimation par pose, avant de calculer la médiane des estimations. Ceci est également valable pour l'étalonnage de la composante translationnelle, qui est théoriquement possible à partir de deux poses du miroir, mais l'estimation sera bien sûr plus robuste si davantage de poses sont utilisées.

En pratique, nous utilisons un minimum de six poses du miroir. Différentes images du miroir, acquises sous différentes poses, sont représentées sur la figure 3, avec une vue de la reconstruction 3D de la scène d'étalonnage.

3. Application : reconstruction 3D par stéréophotométrie

La stéréophotométrie [Woo80] est une technique de reconstruction 3D monoculaire, où la scène est photographiée sous différents éclairages afin d'en extraire le relief et la réflectance. Nous proposons d'utiliser le dispositif multimédia webcam-écran étudié dans le paragraphe précédent comme scanner 3D, à la manière de [FY07, Sch08] : l'écran permet d'afficher successivement m images qui éclairent la scène à reconstruire (cf. figure 4). La webcam capture ces images, desquelles sont extraites hors ligne les caractéristiques géométriques et photométriques de la scène. Ceci est possible grâce au calibrage géométrique du système, qui permet de spécifier géométriquement le flux lumineux incident.

3.1. Modélisation de l'écran vu comme une source lumineuse

Lorsque l'écran est utilisé en mode projection, chacun de ses pixels peut être considéré comme une source lumineuse ponctuelle anisotrope ayant pour direction principale la normale à l'écran [FY07]. Ce dispositif est représenté sur la figure 5.

Soit $\mathcal{E}$ l'ensemble des pixels d'un motif affiché à l'écran, et $\mathbf{X}_\mathcal{E} \in \mathcal{E}$. On suppose que le motif affiché à l'écran est uniforme et que tous les pixels réagissent de la même façon, c'est-à-dire que chaque pixel de $\mathcal{E}$ a la même intensité. En tout point $\mathbf{X} = [x_\mathbf{X}, y_\mathbf{X}, z_\mathbf{X}]^\top$ de l'espace tel que $z > 0$, la contribution du pixel $\mathbf{X}_\mathcal{E}$ à l'éclairage incident en $\mathbf{X}$ peut être modélisée par la formule suivante :

$$\mathbf{L}_{\mathbf{X}_\mathcal{E}}(\mathbf{X}) = \frac{f(\theta)}{\|\mathbf{X} - \mathbf{X}_E\|^3}(\mathbf{X} - \mathbf{X}_\mathcal{E}) \tag{15}$$

où f représente l'anisotropie de la source : f est en général une fonction décroissante de l'angle θ entre la normale à l'écran et la direction $\mathbf{X} - \mathbf{X}_\mathcal{E}$. Cette fonction peut être calibrée par une procédure spécifique [PSM*14], ou choisie empiriquement. Il a été montré dans [Sch08] que les écrans LCD pouvaient être modélisés par une décroissance en cosinus. Grâce au choix de l'écran comme référence du système de coordonnées, cela s'exprime très simplement par $f(\theta) = \frac{z_\mathbf{X}}{\|\mathbf{X} - \mathbf{X}_\mathcal{E}\|}$, ce qui fournit le modèle suivant :

$$\mathbf{L}_{\mathbf{X}_\mathcal{E}}(\mathbf{X}) = \frac{z_\mathbf{X}}{\|\mathbf{X} - \mathbf{X}_\mathcal{E}\|^4}(\mathbf{X} - \mathbf{X}_\mathcal{E}) \tag{16}$$

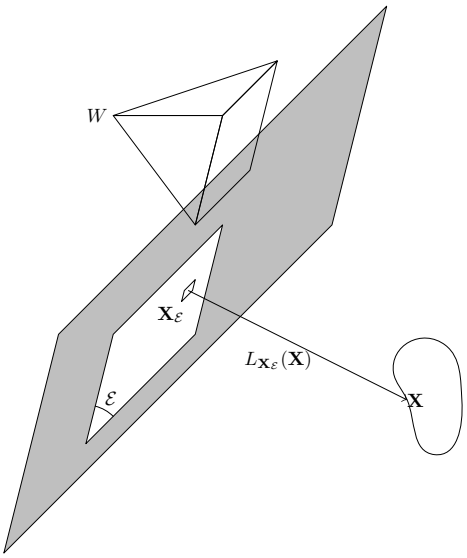


Figure 5: *L'écran utilisé comme source lumineuse. Chaque pixel $\mathbf{X}_\mathcal{E}$ de la partie allumée $\mathcal{E}$ de l'écran émet de la lumière dans toutes les directions. La contribution de ce pixel à l'éclairage reçu par un point $\mathbf{X}$ de la scène est notée $\mathbf{L}_{\mathbf{X}_\mathcal{E}}(\mathbf{X})$.*

3.2. Modélisation photométrique de l'image

Soit $\mathcal{S}$ une surface à reconstruire, placée face au dispositif multimédia webcam-écran. Soit $\mathbf{X} \in \mathcal{S}$ et $\mathbf{N}_\mathbf{X}$ la normale unitaire sortante à la surface au point $\mathbf{X}$.

Comme les différents pixels émettent des ondes lumineuses incohérentes, le faisceau lumineux incident en $\mathbf{X}$ est égal à la somme des contributions des pixels $\mathbf{X}_\mathcal{E}$. D'après le modèle lambertien, la luminance obtenue dans les canaux rouge, vert et bleu, lorsque la scène est éclairée par le motif

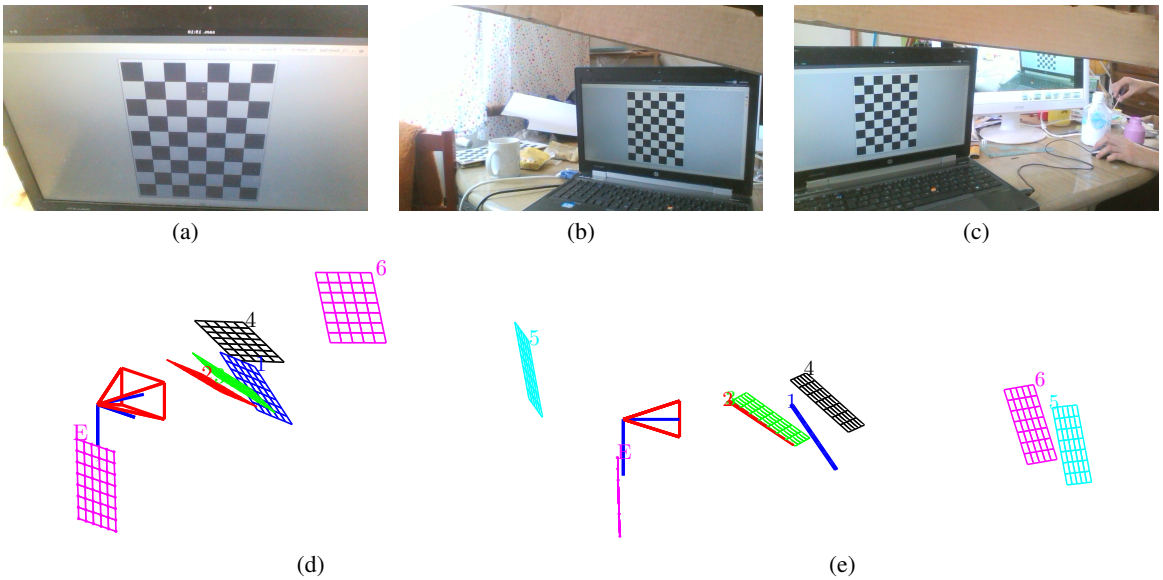


Figure 3: *Résultats de la procédure d'étalonnage appliquée à un ordinateur portable avec caméra intégrée. (a-b-c) Trois des images d'étalonnage acquises par la webcam. La mire est directement affichée à l'écran, et réfléchie par le miroir vers la webcam (le miroir est partiellement visible sur l'image de droite). (d-e) « Scène » reconstruite : les poses des différents miroirs et de la webcam est estimées relativement à l'écran, ce qui permet de spécifier géométriquement l'ensemble de la scène d'étalonnage (les images (a-b-c) correspondent aux poses 1, 5 et 6 de la reconstruction).*

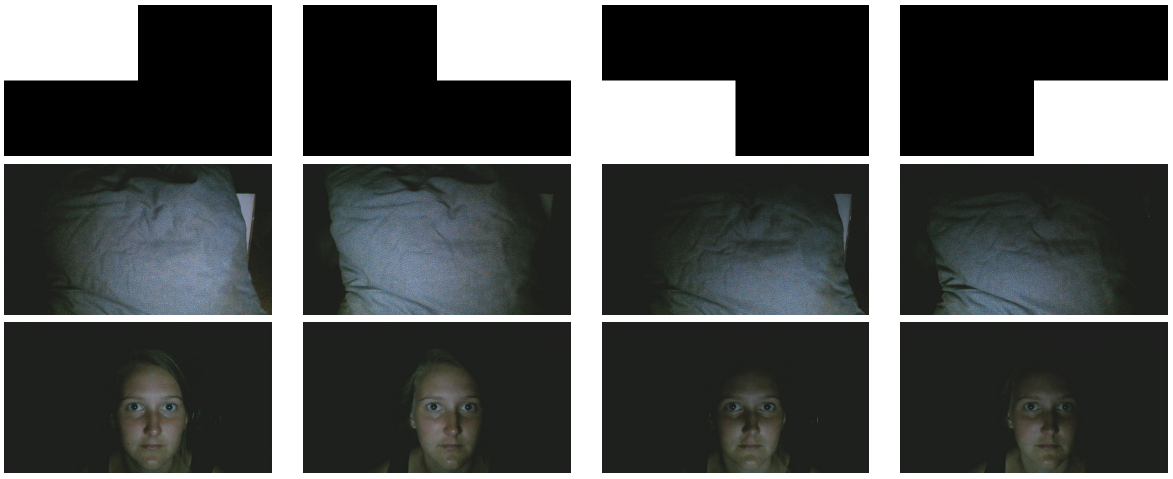


Figure 4: *Le dispositif multimédia utilisé comme un scanner 3D : m différents motifs sont affichés successivement à l'écran (première ligne), ce qui permet d'obtenir m images sous différents éclairages (deuxième et troisième lignes). À partir de ces images et de la pose de la caméra relativement à l'écran, on peut estimer un modèle 3D de la surface.*

$\mathcal{E}_i$, vaut :

$$I^i_{\mathcal{E}}(\mathsf{P}\mathbf{X}) = -\rho_{\mathbf{X}}\mathbf{N}_{\mathbf{X}}^{\top}\sum_{\mathbf{X}_{\mathcal{E}}\in\mathcal{E}_i}\mathbf{L}_{\mathbf{X}_{\mathcal{E}}}(\mathbf{X}) \quad (17)$$

$$= -z_{\mathbf{X}}\,\rho_{\mathbf{X}}\,\mathbf{N}_{\mathbf{X}}^{\top}\sum_{\mathbf{X}_{\mathcal{E}}\in\mathcal{E}_i}\frac{\mathbf{X}-\mathbf{X}_{\mathcal{E}}}{\|\mathbf{X}-\mathbf{X}_{\mathcal{E}}\|^4} \quad (18)$$

où l'albédo $\rho_{\mathbf{X}}$ représente la proportion de lumière réémise par la surface, relativement au canal de couleur considéré. Le problème de la reconstruction 3D par stéréophotométrie à partir du dispositif webcam-écran peut alors être formalisé de la façon suivante :

Problème 3 *À partir de m images I^i, $i = 1\ldots m$, acquises sous m motifs d'éclairage connus $\mathcal{E}_i$, $i = 1\ldots m$, et de la pose* P *de la webcam, estimer les points 3D* $\mathbf{X}$ *de la surface dans le repère de l'écran, ainsi que l'albédo associé* $\rho_{\mathbf{X}}$.

3.3. Inversion du modèle

Afin de résoudre le problème 3, il faut inverser le système formé par les *m* équations (18) correspondant aux *m* prises de vue. Cette étape sera détaillée dans un travail futur dédié au problème de la reconstruction 3D par stéréophotométrie en présence de sources lumineuses « étendues », et ne sera

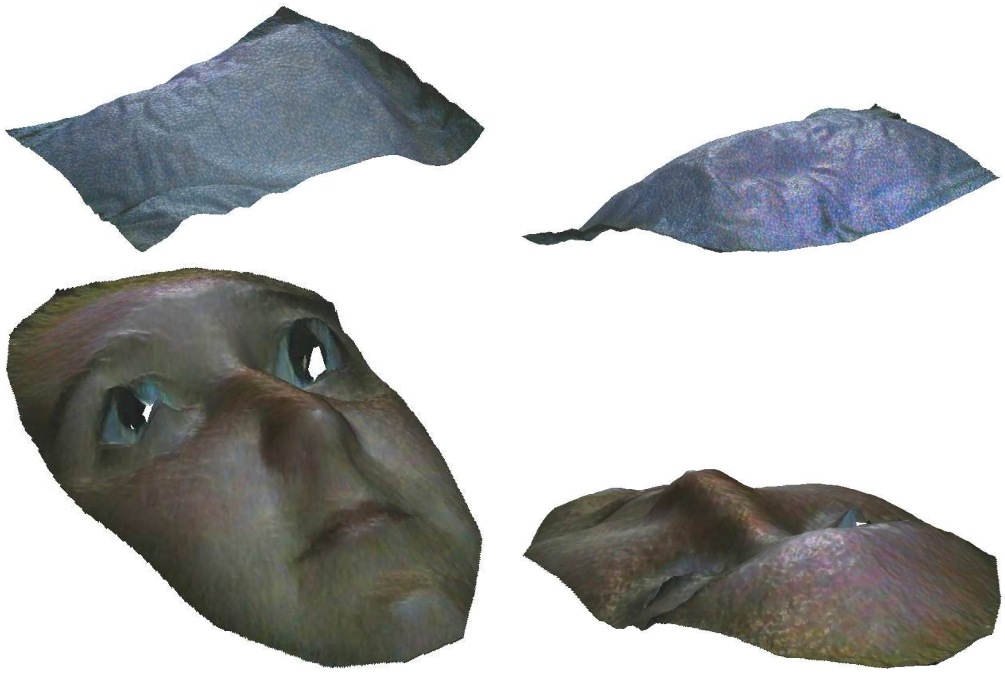


Figure 6: *Résultat de la reconstruction 3D à partir du dispositif multimédia, appliquée aux deux séries de 4 images couleur de la figure 4. Le dispositif permet de retrouver à la fois les coordonnées 3D des points de la surface photographiée, et l'albédo relativement à chacun des canaux de couleur. Le dispositif multimédia utilisé est un ordinateur personnel HP EliteBook 8570w avec caméra HD* 1280×720*.*

donc pas détaillée dans cet article. Notons simplement que, grâce à la présence du terme d'anisotropie, les points $\mathbf{X}$ apparaissent explicitement dans le modèle photométrique de l'image, ce qui n'est généralement pas le cas en stéréophotométrie. En conséquence, l'ambiguïté d'échelle liée à l'intégration perspective du champ de normales est évitée, et il est possible d'obtenir la reconstruction 3D sans ambiguïté. Deux exemples de reconstruction 3D, obtenues à partir des jeux d'images de la figure 4, sont présentés sur la figure 6.

4. Conclusion

Dans cet article, nous avons montré qu'il était possible de transformer n'importe quel dispositif multimédia comprenant un écran et une webcam en un scanner 3D capable d'estimer sans ambiguïté le relief de la scène ainsi que l'albédo relativement à chaque canal de couleur.

Pour ce faire, nous avons d'abord introduit une procédure d'étalonnage géométrique spécifique permettant de retrouver la pose de la webcam, relativement à l'écran qui est utilisé pour afficher une mire d'étalonnage, laquelle mire est réfléchie vers la webcam au moyen d'un miroir plan. Ensuite, en changeant le motif affiché à l'écran, nous avons montré que le dispositif pouvait être utilisé pour effectuer la reconstruction 3D par stéréophotométrie.

Ce travail préliminaire sera complété ultérieurement par une analyse plus profonde du modèle photométrique de l'image. Des évaluations quantitatives des deux étapes (étalonnage et reconstruction 3D) doivent également être effectuées.

Références

[Bou04] BOUGUET J.-Y. : Camera calibration toolbox for matlab.

[FY07] FUNK N., YANG Y.-H. : Using a raster display for photometric stereo. In *Computer and Robot Vision, 2007. CRV '07. Fourth Canadian Conference on* (2007), pp. 201–207.

[HZ03] HARTLEY R., ZISSERMAN A. : *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[PSM*14] PARK J., SINHA S., MATSUSHITA Y., TAI Y., KWON I. : Calibrating a non-isotropic near point light source using a plane. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* (2014), p. To appear.

[RBN10] RODRIGUES R., BARRETO J., NUNES U. : Camera pose estimation using images of planar mirror reflections. In *European Conference on Computer Vision (CVPR 2010)* (2010).

[Sch08] SCHINDLER G. : Photometric stereo via computer screen lighting for real-time surface reconstruction. In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)* (2008).

[Woo80] WOODHAM R. J. : Photometric method for determining surface orientation from multiple images. *Optical Engineering. Vol. 19*, Num. 1 (1980), 139–144.

Reconstruction semi-régulière de surfaces par stéréoscopie

J-L. Peyrot[1], Frédéric Payan [1] et Marc Antonini [1]

[1]Laboratoire I3S - UMR 7271 - Université de Nice - Sophia Antipolis - CNRS
peyrot@i3s.unice.fr, fpayan@i3s.unice.fr, am@i3s.unice.fr

Résumé

Notre objectif consiste à inclure dans les systèmes stéréoscopiques un remailleur semi-régulier qui est capable de générer un maillage semi-régulier à partir des images stéréoscopiques, au contraire des systèmes actuels qui génèrent seulement des nuages de points. Notre méthode de reconstruction est basée sur une approche coarse-to-fine*, et crée directement à partir des images stéréoscopiques une maille semi-régulière. De plus, pour tenir compte des contraintes temps-réel des systèmes d'acquisition, cette construction semi-régulière est parallélisée sur GPU. Les résultats expérimentaux montrent l'efficacité de notre méthode sur divers types de surfaces.*

Mots clé : Maillage semi-régulier, multi-résolution, stéréoscopie, GPU

1. Introduction

Parmi les systèmes de numérisation, les systèmes stéréoscopiques utilisent la lumière comme grandeur physique pour mesurer les informations 3D d'une scène. Ils se composent de deux caméras pour générer deux images correspondant à deux points de vues différents de la même scène, comme l'illustre l'exemple donné à la Figure 1.

Trois étapes sont nécessaires pour reconstruire l'information tridimensionnelle d'une scène :

1. la **calibration** des deux caméras dont le but est de trouver la relation entre les coordonnées spatiales d'un point de la scène et ses coordonnées dans le repère associé aux images stéréoscopiques. Cela nécessite d'estimer les paramètres extrinsèques (*i.e.* la position et l'orientation des caméras) et intrinsèques (*i.e.* la distance focale des caméras, les facteurs d'agrandissement de l'image, les coordonnées de la projection du centre optique des caméras sur les plans images, le facteur de non-orthogonalité qui indique si la grille des cellules photosensibles qui composent le capteur, est rectangulaire ou non). Mathématiquement, on peut ainsi déterminer les matrices de passage du repère de la scène (X_w, Y_w, Z_w) au repère image (U_i, V_i) ;
2. l'**appariement** des deux points de vues, pour trouver les paires de pixels correspondants aux images du même point 3D à travers les deux caméras. Un état de l'art sur les différentes techniques de mise en correspondance peut être trouvé dans la thèse de Chambon [Cha05] ;

Figure 1: *Exemple d'un système de reconstruction par stéréoscopie.*

3. la **triangulation** consistant à calculer les coordonnées 3D des points de la scène à partir des paramètres extrinsèques et intrinsèques déterminés lors de la calibration, et du résultat de l'appariement.

Parmi l'état de l'art sur les méthodes de reconstruction de surfaces, on s'attachera plus particulièrement à décrire les principales techniques de stéréoscopie qui utilisent une paramétrisation lors de la reconstruction. Pour avoir plus de détails et une vue plus globale des techniques de reconstruction de surfaces, nous proposons au lecteur de regarder l'état de l'art proposé par Seitz *et al.* [SCD*06].

Une première méthode que nous pouvons citer est la technique proposée par Park *et al.* [PSM*13] qui réalise une reconstruction efficace à partir d'images multi-vues d'une scène ou d'un objet. L'idée principale consiste à combiner une méthode de reconstruction de type *Multi-View Stereo (MVS)* qui utilise une correspondance entre les pixels des images multi-vues, et une méthode utilisant la notion de reflectance de la surface *Shape from Shading* comme par exemple celle de Zhang *et al.* [ZPKA02]. Pour cela, les auteurs utilisent une structure composée de deux caméras, d'un ensemble de lumières et d'une table pivotante sur laquelle

repose l'objet scanné. Comme le montre la Figure 2, la table pivotante permet d'obtenir plusieurs images de l'objet à différents angles de vue, tandis que l'ensemble de lumière dont une seule est allumée à un instant donné permet de générer plusieurs niveaux d'éclairage de la même scène.

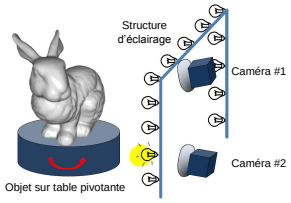


Figure 2: *Structure utilisée pour acquérir des images (image extraite de [PSM* 13]).*

La première étape de cette méthode consiste à générer un maillage de base de l'objet scanné. Pour cela, la technique *Structure from Motion* [SSS06] est tout d'abord utilisée pour générer le nuage de points de l'objet reconstruit. Ensuite une carte de profondeur est créée grâce à la méthode *MVS* de Hernandez *et al.* [HVC08] et utilisée pour construire le maillage de base. La seconde étape est la création d'une paramétrisation par *charts* qui consiste à paramétriser avec le moins de distorsion possible la surface du maillage de base en la partitionnant, comme expliqué dans la méthode de Zhou *et al.* [ZSGS04] : une illustration de la paramétrisation obtenue est donnée à la Figure 3.

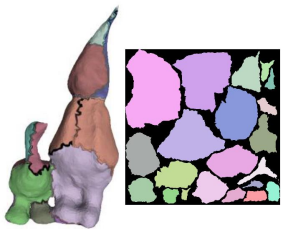


Figure 3: *Illustration des charts définis sur la surface du maillage de base ainsi que la paramétrisation obtenue (image extraite de [PSM* 13]).*

Finalement, à partir de la paramétrisation 2D et d'une estimation des normales à la surface en chaque point du maillage de base, un raffinement est réalisé sur l'espace paramétrique 2D en utilisant une minimisation d'énergie pour contraindre la position des points dans l'espace 3D. Ainsi, cette méthode génère des reconstructions de haute qualité en fusionnant les avantages des méthodes géométriques et photométriques, grâce à l'utilisation d'une paramétrisation globale de la surface de l'objet.

Une autre technique proposée par Pietroni *et al.* [PTSZ11] permet de générer des paramétrisations globales applicables à différents types de surfaces, comme les surfaces implicites, les maillages polygonaux, les nuages de points générés par des scanners 3D, etc. Les auteurs présentent une application de cette technique de paramétrisation au remaillage quadrangulaire de surfaces. Le principe de cette méthode, schématisé à la Figure 4 consiste à paramétriser chaque image de distances (nuage de points d'une vue stéréoscopique par exemple) vers un domaine paramétrique 2D, en tenant compte des artefacts aux frontières des différentes images. En effet, lors de la paramétrisation d'une surface vers un domaine planaire paramétrique, un ensemble de disques topologiques pris sur la surface sont paramétrisés indépendamment, ce qui génère des duplications pour les points de la surface se trouvant aux frontières des disques topologiques car ils appartiennent à plusieurs disques par définition. De ce fait, pour générer une paramétrisation globale robuste, il convient de gérer ces zones de frontières et d'imposer des conditions par rapport aux changements de coordonnées paramétriques lors du passage d'un disque à l'autre. Pour tenir compte des quatre orientations possibles d'un référentiel donné et des différentes positions de l'origine des repères locaux sur le domaine paramétrique, il faut imposer que ce changement de coordonnées paramétriques se fasse suivant une rotation d'angle $\frac{k\pi}{2}$ et une translation de valeur entière. Pietroni *et al.* [PTSZ11] calculent cette paramétrisation en minimisant une contrainte sur les gradients avec l'équation (1) ci-dessous,

$$E(q) = \min_{q} \sum_{T} A_T . ||\Delta q^T - w^T||^2 \qquad (1)$$

avec T un triangle d'une image de distances (les images de distances ont été triangulées), Δq est le gradient au point de coordonnées paramétriques complexe $q = u + \sqrt{-1}v$, et w le vecteur complexe représentant les deux directions de gradient.

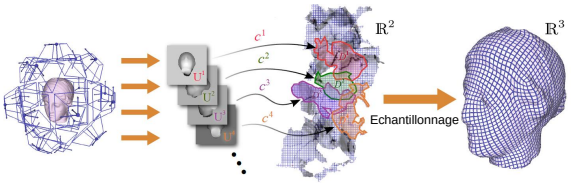


Figure 4: *Principe de la méthode [PTSZ11]. Un ensemble d'images de distances (nuages de points par exemple) U^i est généré. Ensuite, chaque image U^i est paramétrisée dans un domaine planaire 2D grâce à la bijection c^i, en tenant compte des contraintes de distorsion pour générer au final une paramétrisation globale. A droite est présenté un résultat de remaillage quadrangulaire (image extraite de [PTSZ11]).*

Les deux techniques présentées précédemment sont les principales techniques qui utilisent une paramétrisation durant une étape de reconstruction de surfaces. Elles sont fiables et génèrent des maillages de haute qualité en termes de forme et d'aire des triangles (quadrangles).

2. Maillage semi-régulier basé stéréoscopie

2.1. Présentation des maillages semi-réguliers

Un maillage semi-régulier possède une connectivité intéressante car elle permet de générer des maillages à différents

niveaux de détails géométriques, et est très utilisée en compression. En notant L le nombre de résolutions que l'on souhaite créer, on peut générer $L-1$ maillages, indicés de M_0 à M_{L-2} correspondant aux différents niveaux de détails, avec en plus $M_{L-1} = M_{sr}$ car ce dernier correspond au maillage de niveau de détails le plus élevé. De façon générale, à partir du maillage M_l de résolution $l \in [1, L-1]$, on génère le maillage de résolution inférieure M_{l-1} par fusions quaternaires de ses triangles, comme illustré à la Figure 5 : le maillage M_{l-1} correspond donc à une approximation géométrique du maillage de résolution supérieure M_l. La hiérarchie de maillages est créée est réalisant successivement cette étape de fusions quaternaires à partir du maillage semi-régulier $M_{sr} = M_{L-1}$ jusqu'à obtenir le maillage de résolution la plus basse M_0. La Figure 6 montre un exemple d'une telle hiérarchie générée sur le modèle RABBIT à $L = 3$ niveaux de résolution.

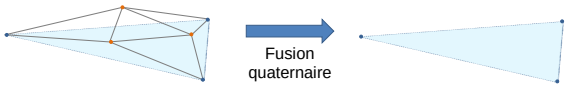


Figure 5: *Fusion quaternaire d'un triangle du maillage M_l de résolution l. A gauche sont représentés 4 triangles voisins du maillage M_l, avec le triangle bleu correspondant au triangle qui sera généré après leur fusion et qui fera partie du maillage de résolution inférieure M_{l-1}, comme illustré à droite.*

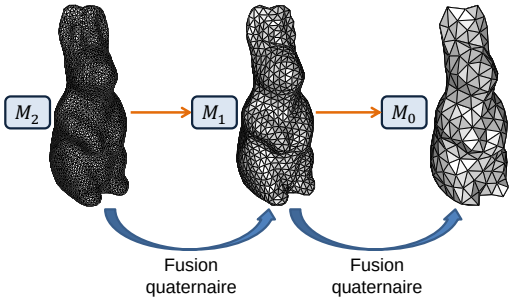
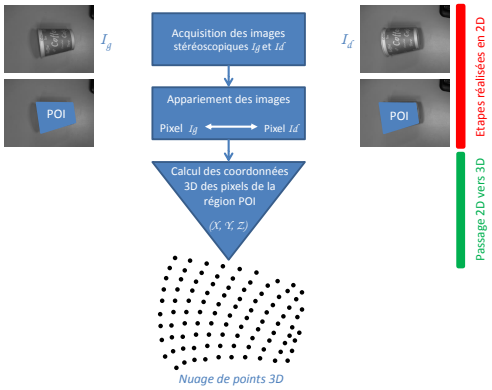


Figure 6: *Hiérarchie multi-résolution du modèle* RABBIT *avec 3 niveaux de détails.*

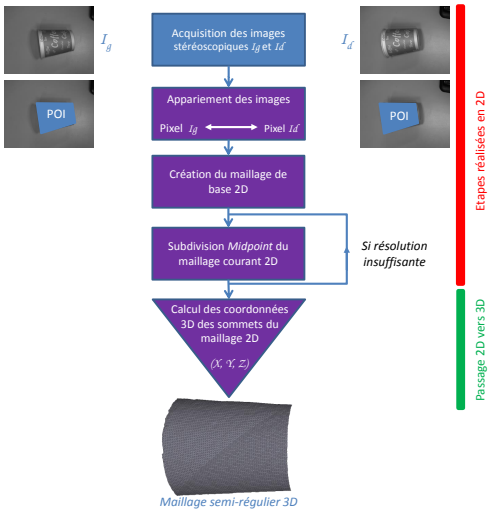
2.2. Méthode de maillage semi-régulier proposée

La Figure 7(a) illustre les deux principales étapes d'un algorithme classique de reconstruction stéréoscopique.

1. Tout d'abord, une méthode d'appariement [SS02] recherche les paires de pixels dans les deux images stéréoscopiques, qui représentent le même point 3D dans la scène. L'ensemble des paires de pixels est nommé région *Pixels Of Interest (POI)*, correspondant aux régions entourées en bleu sur les images du haut de la Figure 7 ;
2. Chaque paire de pixels dans la région POI est ensuite utilisée pour calculer les coordonnées du point 3D correspondant dans la scène, à l'aide d'une technique de triangulation [HZ04].



(a) Approche classique.



(b) Approche proposée.

Figure 7: *Approche classique (en haut) et l'approche proposée (en bas) pour reconstruire une surface à partir d'un système stéréoscopique.*

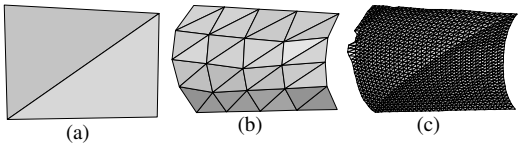
La sortie des scanners 3D actuels consiste en un ensemble de points 3D, appelé communément un nuage de points 3D. Ce nuage de points est généralement dense, puisqu'il contient autant de points 3D que de paires de pixels dans la région POI, et plus la résolution des images stéréoscopiques est élevée, plus le nuage de points sera dense.

Notre approche est différente comme le décrit la Figure 7(b), et est constituée de quatre étapes :

1. Détecter la région POI dans les images stéréoscopiques grâce à la phase d'appariement ;
2. Créer le maillage de base 2D à partir de la région POI ;
3. Subdivision de la connectivité du maillage 2D courant et repositionnement des sommets ajoutés, à l'intérieur de la région POI ; cette phase est réitérée jusqu'à obtenir un maillage à la résolution souhaitée ;
4. Calcul des coordonnées 3D des sommets du maillage 2D pour générer le maillage semi-régulier 3D final.

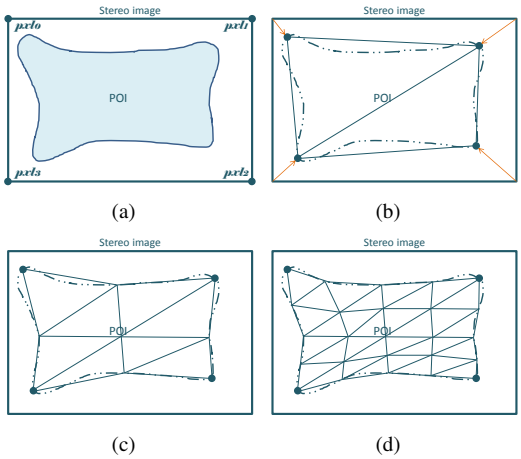
Notre algorithme de maillage a l'avantage de ne nécessiter comme entrées que les images stéréoscopiques et la connaissance de la région POI pour générer la connectivité semi-régulière. De plus, au lieu de calculer les coordonnées 3D pour chaque pixel de la région POI, comme cela est fait avec des algorithmes classiques (ce qui génère des nuages de points sur-échantillonnés), notre méthode restreint les calculs aux seuls pixels utilisés lors des subdivisions successives, ce qui diminue la complexité par rapport aux approches classiques.

Un exemple est présenté à la Figure 8 sur un maillage semi-régulier à 3 niveaux de résolution généré par notre méthode.



(a) (b) (c)

Figure 8: *Trois niveaux de résolution d'un maillage semi-régulier créé avec notre algorithme de reconstruction semi-régulier. Chaque résolution contient (a) 2 triangles ; (b) 32 triangles et (c) 2048 triangles.*



(a) (b) (c) (d)

Figure 9: *Génération* coarse-to-fine *du maillage semi-régulier* 2*D. (a) Phase d'initialisation : quatre sommets sont positionnés aux coins de l'image stéréoscopique ; (b) création du maillage de base ; (c) et (d) les deux premières résolutions obtenues par subdivisions* Midpoint *après déplacements des sommets ajoutés sur la région POI à l'aide de l'algorithme de plus proches voisins parallélisé sur GPU.*

2.3. Détails d'implémentation

Génération du maillage semi-régulier 2D

Comme expliqué précédemment, notre algorithme crée d'abord la connectivité du maillage semi-régulier dans l'espace 2D des images stéréoscopiques. Pour générer cette connectivité semi-régulière, nous adoptons une approche *coarse-to-fine*. On crée d'abord un maillage de base qui est une version très approximative de la région POI, en utilisant les trois phases suivantes, illustrées sur les Figures 9(a) et 9(b) :

1. Quatre points sont initialisés aux coins de l'image stéréoscopique ;
2. Un algorithme parallélisé sur GPU de plus proches voisins déplace ces points sur leur plus proches pixels dans la région POI (flèches oranges dans la Figure 9(b)) ;
3. La connectivité des deux triangles correspondant est créée.

La connectivité des résolutions supérieures est générée avec une subdivision *Midpoint* : chaque arête est divisée en deux arêtes plus petites en ajoutant un sommet en son milieu et chaque triangle est ainsi divisé en quatre sous-triangles. Un tel maillage est appelé semi-régulier car les sommets ajoutés sont toujours réguliers (leur valence est égale à 6). Le point milieu ne faisant pas forcément parti de la région POI, il est déplacé vers le plus proche pixel de la région POI grâce à l'algorithme de plus proches voisins.

La Figure 9(c) montre le maillage 2D semi-régulier résultant. Le même processus est réitéré plusieurs fois pour générer des maillages de plus en plus denses et détaillés : la Figure 9(d) montre le troisième niveau de détails de la connectivité définie par $2 \times 4 \times 4$ triangles.

Calculs des coordonnées 3D

Pour générer le maillage final, les coordonnées 3D de chaque point du maillage semi-régulier 2D sont calculées. Nous utilisons la même technique que l'approche classique, décrite à la Figure 7(a). En itérant cette technique sur chaque sommet du maillage semi-régulier 2D, on crée le maillage semi-régulier 3D, comme illustré sur la Figure 8. Nous observons que le maillage semi-régulier résultant approxime fidèlement la surface originale, en évitant la génération d'un maillage dense : le nuage de points contient 93595 points 3D, alors que la résolution la plus fine de notre maillage semi-régulier ne contient que 4225 sommets.

3. Résultats expérimentaux

Nous présentons plusieurs résultats pour montrer l'efficacité de notre méthode, sur divers types de surfaces. La Figure 10 présente une image stéréoscopique d'un coin de mur, en dessous sont présentés les mailles semi-régulières à différents niveaux de résolution obtenues avec notre mailleur.

La Figure 11 présente le maillage semi-régulier de plus haute résolution obtenu pour une partie d'une porte (à gauche) et sa représentation texturée (à droite).

4. Conclusion

Nous avons présenté une contribution permettant de construire directement à partir des images stéréoscopiques une maille semi-régulière, sans passer par l'information 3D, au contraire des méthodes actuelles. Cela permet ainsi de raccourcir la chaîne de numérisation classique qui consiste à partir du nuage de points généré par le scanner, à le nettoyer, le mailler puis ensuite à le remailler en semi-régulier.

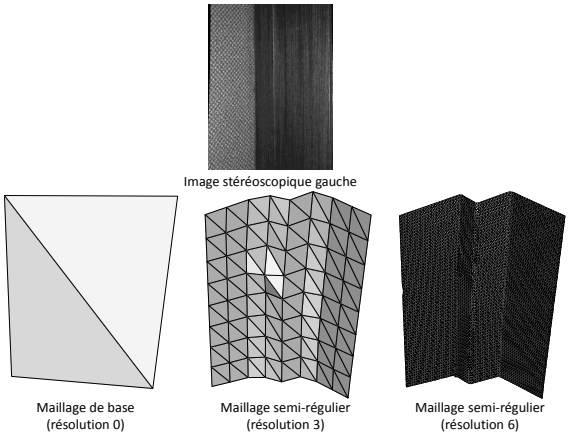


Figure 10: *Résultat visuel d'un coin de mur.*

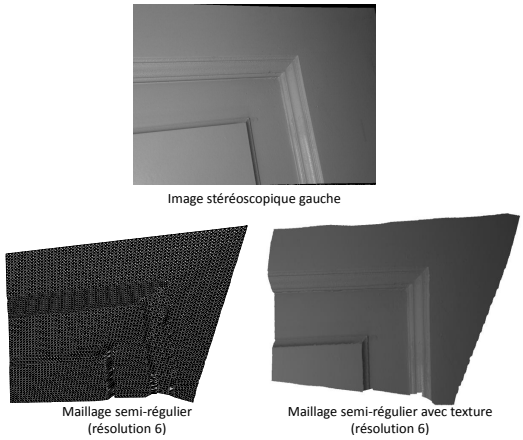


Figure 11: *Résultat visuel d'une partie d'une porte.*

Une perspective possible consiste à améliorer notre mailleur semi-régulier pour générer des mailles plus uniformes en termes de forme et d'aire des triangles. Ceci permettrait de régulariser les distributions de sommets, ce qui facilite l'étape de subdivision et évite la génération de triangles fins et disproportionnés.

Références

[Cha05] CHAMBON S. : *Mise en correspondance stéréoscopique d'images couleur en présence d'occultations*. PhD thesis, Université toulouse III-Paul Sabatier, 2005.

[HVC08] HERNANDEZ E. C., VOGIATZIS G., CIPOLLA R. : Multiview photometric stereo. *IEEE Trans. Pattern Anal. Mach. Intell.. Vol. 30*, Num. 3 (mars 2008), 548–554.

[HZ04] HARTLEY R. I., ZISSERMAN A. : *Multiple View Geometry in Computer Vision*, second ed. Cambridge University Press, ISBN : 0521540518, 2004.

[PSM*13] PARK J., SINHA S. N., MATSUSHITA Y., TAI Y.-W., KWEON I. S. : Multiview photometric stereo using planar mesh parameterization. *Computer Vision, IEEE International Conference on. Vol. 0* (2013), 1161–1168.

[PTSZ11] PIETRONI N., TARINI M., SORKINE O., ZORIN D. : Global parametrization of range image sets. *ACM Trans. Graph.. Vol. 30*, Num. 6 (décembre 2011), 149 :1–149 :10.

[SCD*06] SEITZ S. M., CURLESS B., DIEBEL J., SCHARSTEIN D., SZELISKI R. : A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1* (Washington, DC, USA, 2006), CVPR '06, IEEE Computer Society, pp. 519–528.

[SS02] SCHARSTEIN D., SZELISKI R. : A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision. Vol. 47*, Num. 1-3 (avril 2002), 7–42.

[SSS06] SNAVELY N., SEITZ S. M., SZELISKI R. : Photo tourism : Exploring photo collections in 3d. *ACM Trans. Graph.. Vol. 25*, Num. 3 (juillet 2006), 835–846.

[ZPKA02] ZHANG Y., PAIK J., KOSCHAN A., ABIDI M. A. : A simple and efficient algorithm for part decomposition of 3-d triangulated models based on curvature analysis. In *in Proceedings of the International Conference on Image Processing, III* (2002), pp. 273–276.

[ZSGS04] ZHOU K., SYNDER J., GUO B., SHUM H.-Y. : Iso-charts : Stretch-driven mesh parameterization using spectral analysis. Eurographics.

Détection des yeux, du nez et de la bouche par filtres de Haar adaptatifs

N.Pyun[1,2] , M. Marmouget[2] et N. Vincent[1]

[1]Université de Paris Descartes, Paris, France
[2]Konbini, Paris, France

Résumé

L'extraction des yeux, du nez et de la bouche du visage humain sont des tâches largement étudiées dans le domaine de la reconnaissance de formes. Localiser ces régions anatomiques pertinentes du visage est souvent la première étape de nombreuses approches de la vision par ordinateur, comme la segmentation, la reconnaissance ou l'identification de personne, la reconnaissance de l'expression ou de l'émotion du visage, la localisation de points d'intérêts, l'estimation de pose ou encore le suivi du visage. La télésurveillance, l'indexation automatique ou semi-automatique d'images ou de vidéos, la robotique sont autant de domaines applicatifs. Dans cet article, nous proposons une méthode basée sur l'analyse des lignes horizontales. Elles sont extraites d'une carte d'énergie calculée sur des filtres de Haar adaptatifs. L'introduction de connaissances, notamment sur les positions des différentes régions anatomiques pertinentes, ainsi que sur leurs relations spatiales nous permet de les séparer. Une des difficultés majeures de la détection des éléments anatomiques pertinents du visage réside dans la variabilité de l'illumination d'un visage à l'autre, mais aussi des conditions d'illumination inégale sur un visage donné. Afin de rendre la méthode robuste à ces variations d'illumination, nous proposons une analyse multi-seuils capable de choisir, pour chaque région anatomique, un seuil adéquat sur la carte d'énergie horizontale. Notre approche est testée sur les bases BioID, Color FERET et LFW et montre des résultats prometteurs.

Extracting human eyes, nose and mouth are widely studied tasks in pattern recognition. Finding localization of these relevant face anatomic regions is often the first step in many approaches in computer vision, such as segmentation, person recognition or identification, facial expression or emotion recognition, landmarks localization, head pose estimation or face tracking. Such methods are used in many applications such as telemonitoring, automatic and semi-automatic indexation of images and videos, or in robotics. In this paper, a method based on horizontal lines analysis is proposed. Lines are extracted from a energy map computed on adaptive Haar-like features. Briging knowledge, in particular, related to positions of these relevant facial anatomic regions, as well as their relative positions enable to separate them. One of the difficulties of detecting these components lies in illumination variability of a face compared to another, as well as inegal illumination on a single given face. To overcome these illumination variations which often occur, we propose a multi-thresholds analysis able to choose, for each anatomic region, a suitable threshold of the energy map. Our approach is tested on BioID, Color FERET and LFW databases and shows promising results.

Mots clé : Œil, yeux, nez, bouche, Haar, carte d'énergie, analyse multi-seuils, relations spatiales, connaissance

1. Introduction

Extraire les caractéristiques du visage est une étape nécessaire dans de nombreuses applications, comme la reconnaissance de visage [JP09], l'estimation de pose [MCT09], le suivi du visage [MZW10] ou encore la reconnaissance de l'expression faciale [LZM12]. En particulier, les positions des yeux, du nez et de la bouche sont des informations souvent recherchées. Par exemple, dans le suivi du visage, de nombreuses méthodes recourent aux AAM (Active Appareance Models) qui font preuve d'efficacité et de précision. La première étape des AAM [TCT01] implique l'apprentissage de la déformation des visages en appliquant une analyse en composante principale sur les positions et intensités de points saillants (du contours des yeux, du nez et de la bouche) du visage. La seconde étape consiste à la mise en correspondance d'un jeu de points avec le modèle issu de l'apprentissage. Le principal inconvénient des AAM réside dans la nécessité de placer manuellement ces points lors de la phase d'apprentissage. Bien que trouver les boîtes englobantes des yeux, du nez et de la bouche soit insuffisant pour

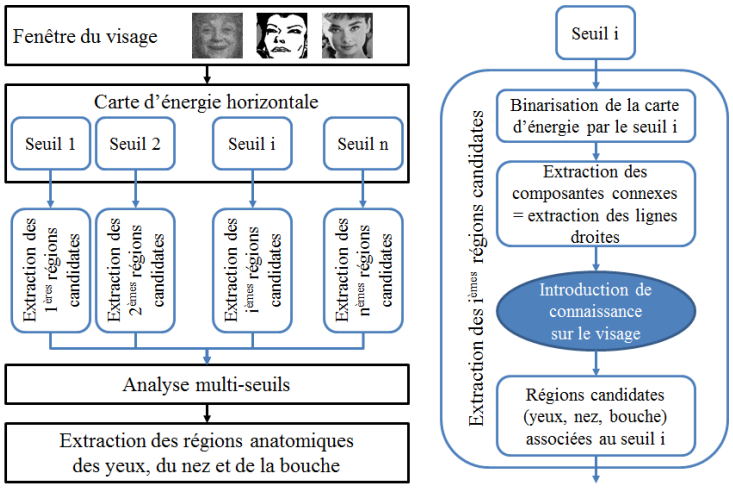


Figure 1: *Schéma global de la méthode.*

déterminer les points saillants, cela permettrait d'améliorer la précision, puisque ces boîtes englobantes réduisent la zone de recherche.

Il existe de nombreuses méthodes capable d'extraire des caractéristiques à partir des visages. Elles peuvent être regrouper en trois catégories. La première est constituée des approches basées sur l'apparence [IC13], [LR12], [qZhC12]. Dans cet ensemble, le but est d'extraire des caractéristiques locales dans un sous-espace approprié en faisant appel aux techniques d'apprentissage. Par exemple, dans [VP05], GentleBoost est utilisé sur des filtres de Gabor locaux ou encore dans [AB10], l'analyse en composantes principales est appliquée sur les LTP [TT10]. Les AAM font partie de cette catégorie. Ces méthodes nécessitent une étape d'apprentissage, et donc une base conséquente d'images annotées. Cependant, la littérature montre qu'elles parviennent à obtenir de bons résultats.

La seconde catégorie est constituée d'approches où un template est utilisé. Dans ces méthodes [DZZ14], la première partie consiste à définir un template spécifique d'une partie du visage (par exemple les yeux). Généralement, ce template contient les contours et cherche à modéliser les possibles déformations. Puis, différents candidats sont extraits avant de les comparer au template. Dans [JH09], les auteurs extraient les yeux en utilisant un template multi-angle. Les candidats sont extraits à l'aide d'opérations morphologiques et des informations sur la symétrie des yeux. Dans [AYH89], un template déformable est utilisé. Celui-ci se déforme en minimisant un coût afin de trouver la meilleure correspondance. Le principal inconvénient de ces méthodes réside dans la difficulté à les généraliser lorsque les conditions de vue, d'illumination ou d'échelle changent. Par exemple, un template définissant des yeux ne parvient à en trouver que lorsqu'ils sont ouverts. Par contre, ces approches ont l'avantage d'être capables de trouver ce qu'elles cherchent dans des images où n'apparaitraient que l'élément recherché.

Enfin, la dernière catégorie regroupe des approches qui incluent de la connaissance et des informations spatiales sur le visage [DPM11], [ZZ12] . Dans [GS07], Les points d'intérêt sont détectés automatiquement sur des visages aux expressions variées. Des informations spatiales sont introduites pour améliorer la précision de la localisation de ces points. Dans [KP97], les auteurs proposent une méthode de détection de visage qui requiert des règles relatives aux informations spatiales du visage.

La méthode que nous proposons appartient à cette dernière catégorie. Le but est d'extraire les rectangles englobants des yeux, du nez et de la bouche. Elle inclut de la connaissance sur le visage, notamment sur la distribution spatiale de ces caractéristiques sur le visage. Par exemple, les yeux se situent sur la partie supérieure du visage, le nez et la bouche sont alignés sur l'axe de symétrie du visage, etc. Afin de parvenir à de bonnes détections et malgré la variation d'échelle, une carte d'énergie s'adaptant à l'échelle du visage est proposée. Les connaissances anatomiques sur les tailles des différents éléments recherchés, ainsi que leurs positions relatives nous permettent de fixer certaines limites dans nos calculs.

La section suivante présente la méthode de façon générale. la section 3 est consacrée à l'extraction des régions anatomiques candidates. La section 4 décrit l'extraction des régions anatomiques finales grâce à une analyse multi-échelle. La section 5 est consacrée à l'évaluation de la méthode.

2. Architecture générale

2.1. Vue générale de la méthode

Dans la littérature, les éléments anatomiques du visages sont souvent déterminés individuellement ; certains recherchent les yeux, d'autres la bouche, ou plus rarement le nez. Pourtant, il est certain qu'ils peuvent être utiles ensemble, par exemple pour établir un modèle 3D ou encore pour faire de la reconnaissance. C'est donc l'objectif de cet article. Toutefois, nous nous plaçons ici au niveau des boites englobantes et non de points anatomiques. L'avantage de se placer à ce niveau est la possibilité de détecter l'élément anatomique en question, malgré la présence visible de points

anatomiques due à des occlusions ou à des conditions d'illumination inégales. On pourrait rechercher indépendamment chaque élément anatomique, car ils ne sont pas toujours illuminés de la même manière. Toutefois, en les recherchant ensemble, nous pouvons utiliser leurs positions relatives qui constituent une source d'information.

La figure 1 présente une vue globale de la méthode. D'un point de vue général, la méthode proposée est constituée de trois étapes principales. Tout d'abord, à partir des fenêtres de visage, une carte d'énergie horizontale est calculée. Dans un second temps, nous cherchons à extraire des régions anatomiques candidates. En effet, les conditions d'illuminations peuvent varier d'un élément anatomique à l'autre, un seuil global, appliqué à la carte, serait alors insuffisant. A ce stade, nous avons donc quatre ensembles distincts regroupant respectivement les candidats de l'œil droit et gauche, du bout du nez et de la bouche. La dernière étape consiste à choisir, pour une région anatomique donnée, un candidat par une analyse multi-seuil. Comme nous venons de voir, la deuxième étape consiste à extraire les régions anatomiques candidates. Pour un seuil de binarisation donné de la carte d'énergie horizontale, on extrait les composantes connexes (CC), ce qui revient à extraire les lignes horizontales. Puis en introduisant de la connaissance sur le visage, notamment sur les positions et tailles relatives des différents éléments anatomiques, nous parvenons à extraire les boites englobantes candidates de l'œil droit, gauche, du nez et de la bouche associées à un seuil de binarisation.

2.2. Carte d'énergie horizontale

Tout d'abord, les visages sont extraits dans des fenêtres rectangulaires, par exemple en utilisant la méthode de Viola et Jones [VJ01] ou encore les LBP [TOM02]. Nous supposons que la détection du visage réussit ; le visage est détecté dans son intégralité. Ainsi, l'échelle du visage correspond approximativement à la taille de la fenêtre englobante du visage. Les lignes du visage qui correspondent aux régions saillantes sont essentiellement horizontales (voir figure 2). Cette section décrit les étapes qui nous permettent d'extraire les lignes horizontales.

Figure 2: *Détection de lignes verticales et horizontales du visage par convolution.*

De nombreuses approches permettent d'extraire les lignes horizontales. Dans une transformée de Fourier, nous nous intéresserions aux fréquences verticales. Plus récentes, les transformées en ondelettes apportent beaucoup plus d'information tout en conservant une certaine localité à l'information. Les ondelettes de Haar mettent en œuvre des échelles différentes. La difficulté est alors de choisir le bon niveau d'observation et donc le critère qui permet de le déterminer en fonction des coefficients calculés, et aussi de sélectionner les meilleurs coefficients. Grâce à la connaissance de la fenêtre englobant le visage, nous avons une idée de l'échelle des éléments recherchés. Les largeurs et hauteurs de la bouche et des yeux sont du même ordre de grandeur, la taille de la matrice de convolution utilisée peut donc être fixée et ne dépend que de la taille de la fenêtre du visage. La méthode de Viola et Jones recherche la nature et la taille des bons motifs à appliquer lors de la détection. Puisque la nature (lignes horizontales) et l'échelle des éléments recherchés sont connues, la matrice de convolution de notre méthode correspond au motif horizontal décrit par Viola et Jones (Figure 3).

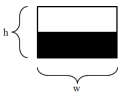


Figure 3: *Filtre de Haar horizontal de hauteur h et de largeur w.*

D'un côté, si le détecteur est trop local, les résultats sont trop bruités. De l'autre, s'il ne l'est pas assez, le détecteur donne des résultats où manquent des informations. Nous utilisons une matrice de convolution horizontale dont la taille dépend seulement de celle de la fenêtre de visage. Soit H et L, la hauteur et largeur respectives de la fenêtre de visage, nous définissons h et l, la hauteur et largeur de la matrice de convolution par la formule (1).

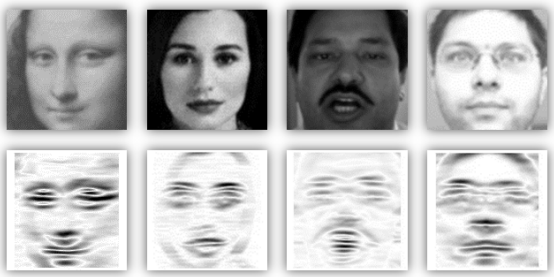


Figure 4: *Les fenêtres de visages sont sur la première ligne, les cartes d'énergie horizontales normalisées E_{norm} sont sur la seconde. plus la valeur est faible, plus le pixel est noir.*

$$\begin{cases} h = max(2; 2 \cdot \alpha) \; et \; l = max(2; 3 \cdot \alpha) \\ avec \; \alpha = min(H/40; L/40) \end{cases} \quad (1)$$

Toutes ces remarques nous conduisent à appliquer à la fenêtre de visage I, un filtre de convolution de noyau H_{lh} (équation (2)).

$$J = I * H_{lh} \quad (2)$$

La carte d'énergie horizontale est alors donnée par l'équation (3).

$$E(X,Y) = |J(X,Y)| \quad (3)$$
$$E(X,Y) = \left| \sum_{blanc} p(x_1,y_1) - \sum_{noir} p(x_2,y_2) \right|$$

Puis, la carte d'énergie est normalisée en E_{norm} par la valeur maximale M de la partie centrale ($a = 1/3 \cdot X$ et $b = 2/3 \cdot X$) de E par l'équation (4) et (5). En effet cette partie centrale ne contient que des pixels issus du visage tandis que les tiers droit ou gauche contiennent souvent des éléments de l'arrière-plan.

$$M = \max_{a<X<b} E(X,Y) \quad (4)$$

$$E_{norm}(X,Y) = 1 - min\left(1, \frac{E(X,Y)}{M}\right) \quad (5)$$

Cette normalisation nous donne l'ordre de grandeur de la variation de la carte d'énergie au niveau du visage. Après cette normalisation, et contrairement à E, plus une valeur dans E_{norm} est faible, plus nous sommes confiant en la présence d'une ligne horizontale dans le voisinage. La Figure 4 montre quelques exemples de cartes d'énergie horizontales normalisées.

3. Extraction des régions anatomiques candidates

Une fois E_{norm} obtenue, nous devons extraire les lignes à direction horizontale. Différents seuils sont appliqués sur E_{norm}. La Figure 5 montre des cartes d'énergie binarisées à différents seuils d'un même visage. Malgré la normalisation de la carte d'énergie, un seuil adéquat global pour tous les visages n'existe pas. De plus, comme l'illumination sur un visage donné peut être irrégulière, un seuil fixe ne peut être appliqué à tous les éléments de ce visage. Pour toutes ces raisons, les régions anatomiques candidates (RAC) des yeux, du nez et de la bouche sont détectés pour chaque seuil. Cette section concerne l'extraction de RAC pour un seuil fixe donné.

Figure 5: *Seuillage de la carte d'énergie. L'image de gauche est la carte d'énergie normalisée, les autres sont les images binaires de cette carte d'énergie (de gauche à droite, les seuils sont de 0,4 ; 0,6 ; 0,8 et 0,99).*

Après le seuillage, les composantes connexes (CC) sont extraites ainsi que leur boîtes englobantes. Le but est alors de regrouper les boîtes représentant respectivement chacun des yeux droit et gauche, le nez et la bouche. Tout d'abord, nous calculons le nombre de pixels appartenant aux CC sur chaque ligne. Toutes les CC qui ont une projection commune sur l'axe des ordonnées sont fusionnées pour former une RAC. Deux RAC consécutives sont fusionnées en une si la distance maximum entre elles est plus petite que $max(1; H/40)$, comme le montre la figure 6.

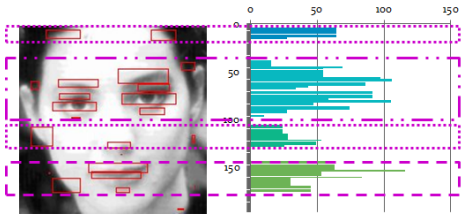


Figure 6: *Histogramme des largeurs des CC sur l'axe des ordonnées. 4 RAC sont extraites de cet histogramme.*

Après cette étape, une liste de RAC rangée est construite. Chaque RAC est représentée par les CC inclues, le point supérieur gauche S et le point inférieur droit T. Cependant, lorsque H ou W est inférieur à 60 pixels, les projections des CCs sur l'axe des ordonnées ne sont plus séparées ; chaque CC devient alors un RAC. Puis, les RAC sont rangées par rapport à l'ordonnée y_S.

3.1. Extraction des RAC des yeux

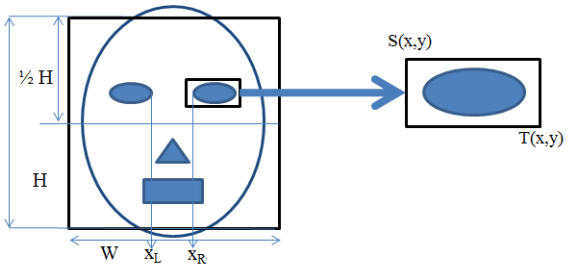


Figure 7: ***Gauche*** *: Connaissances basiques utilisées dans notre méthode.* ***Droite*** *: Exemple d'une RAC définie par le point supérieur gauche S et le point inférieur droit T.*

Afin d'extraire les régions saillantes du visage, des informations sur la distribution spatiale du visage sont utilisées (Figure 7). Puisque les yeux sont situés sur la partie supérieure du visage. Ils sont contenus dans la RAC dont l'ordonnée $y_S{}^*$ respecte les conditions (6).

$$\begin{cases} y_S{}^* < 1/2 \cdot H \\ y_S{}^* = min(|y_S - 1/2 \cdot H|) \end{cases} \quad (6)$$

Les RAC au-dessus de la RAC des yeux sélectionnée et proches du bord supérieur de la fenêtre du visage sont supprimées. Sinon, elles sont fusionnées avec la RAC des yeux. Par exemple dans l'exemple de la Figure 6, seules les deux RAC supérieures sont, dans un premier temps prises en compte. Comme la RAC la plus proche du bord supérieur

est trop éloignée de l'autre, seule la RAC proche du centre de l'image est prise en compte.

Notons qu'à ce moment, seule l'ordonnée et la hauteur des deux sont identifiés dans la RAC des deux yeux.

Puis, les RAC de l'œil droit et gauche sont extraites à partir de la RAC des deux yeux. Les occurrences des boîtes englobantes des CC contenues dans la RAC des deux yeux sont projetées sur l'axe des abscisses. Les rectangles ayant une projection commune sur l'axe des abscisses sont fusionnés pour former une RAC d'un seul œil., comme le montre la figure 8.

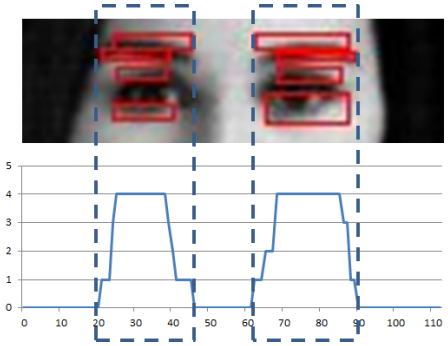


Figure 8: *Projection des occurrences des CC sur l'axe des abscisses de la RAC des deux yeux. les CC avec une projection commune sont fusionnées pour former la RAC de l'œil droit et gauche.*

Deux cas se présentent alors à nous : soit les projections des RAC de l'œil droit et gauche sont séparées, soit elles ne le sont pas. La région des yeux n'est plus une seule RAC, mais une liste de RAC. Puisque les yeux droit et gauche sont les plus significatifs dans la RAC des yeux. Seules les deux RACS ayant les aires les plus élevées sont conservées.

Si la seconde plus grande RAC a une aire ne dépassant pas 10% de celle ayant l'aire la plus élevée, alors elle n'est pas prise en compte. Sinon, la RAC ayant la plus petite abscisse devient la RAC de l'œil droit et l'autre devient la RAC de l'œil gauche.

Souvent, les projections des yeux droit et gauche ne sont pas séparées. Cela arrive lorsque L et H sont basses ou lorsque le seuillage de la carte d'énergie est élevé ou encore lorsque le sujet porte des lunettes. Si, nous n'avons toujours qu'une seule RAC au lieu de deux, afin de séparer la RAC des deux yeux en deux, nous calculons l'histogramme des occurrences des pixels de la carte d'énergie binarisée sur l'axe des abscisses. Bien que les yeux droit et gauche ne soient pas séparés sur cet histogramme, les valeurs entre les deux yeux sont nettement inférieures à celles se trouvant au niveau des yeux. Nous utilisons alors un simple level set supérieur dont la valeur de séparation est de la moitié de la valeur maximale de l'histogramme (Figure 9). Les deux RAC les plus larges sont alors prises en compte et forment les RAC des yeux droit et gauche. Afin de maintenir la cohérence des RAC, les points S, T ainsi que la taille des CC contenues dans chaque RAC détectée, sont modifiées en conséquence.

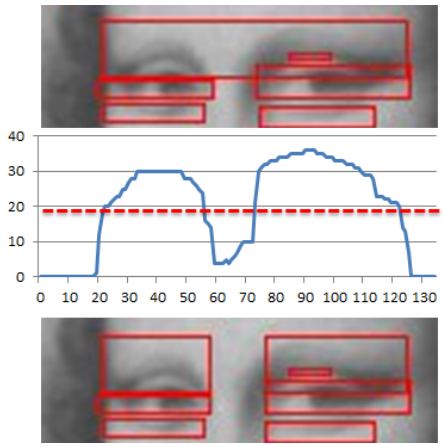


Figure 9: *Projection des pixels se trouvant sur les CC de la RAC des deux yeux. Le level set supérieur permet de séparer la RAC des yeux en RAC de l'œil droit et gauche.*

On note que les RAC de l'œil droit et gauche détectées sont parfois trop larges, car elles incluent les sourcils ou l'arcade sourcilière. D'un autre côté, le level set supérieur permet d'exclure les sourcils ou l'arcade sourcilière par rapport à l'axe des abscisses. C'est pourquoi un level set supérieur est encore une fois appliqué sur chacune des RAC des yeux. Au final, nous obtenons les deux RACS correspondant chacun à l'œil droit et gauche.

3.2. Extraction des RAC du nez et de la bouche

Ici, nous introduisons une nouvelle connaissance commune du visage humain : le nez et la bouche sont situés sur l'axe de symétrie du visage. Cet axe passe par l'espace qui sépare les deux yeux. Ainsi, le nez et la bouche ont une partie située entre les yeux.

A partir des RAC de l'œil droit et gauche, les abscisses x_R et x_L de l'intervalle IGD entre les yeux sont déterminées. La première étape consiste à conserver seulement les boîtes englobantes dont la projection verticale intersecte IGD. Les CC qui ne sont pas inclues dans une des RAC des yeux et dont l'abscisse d'au moins un point de celles-ci est comprise dans $[x_L, x_R]$ sont conservées. En d'autres termes, seules les CC des yeux et celles situées sur l'axe de symétrie verticale du visage sont conservées. Cette étape permet de supprimer les CC situées sur le bord du visage comme le montre la Figure 10.

Puis, afin de séparer la RAC du nez de la RAC de la bouche, une autre connaissance basique sur le visage humain est utilisée. La bouche a une largeur plus importante que celle de la base du nez. Parmi les CC conservées celle qui possède la plus grande largeur (CC_{bouche}) appartient à la bouche. D'un autre côté, la CC qui est la plus proche des yeux (CC_{nez}) appartient au nez.

Toutes les CC en-dessous de CC_{bouche} appartiennent à la bouche. Quant aux CC restantes, elles sont comparées et classifiées en fonction de leur largeur et celles de CC_{bouche} et CC_{nez}. A partir de ces deux ensembles de CC, celui qui se

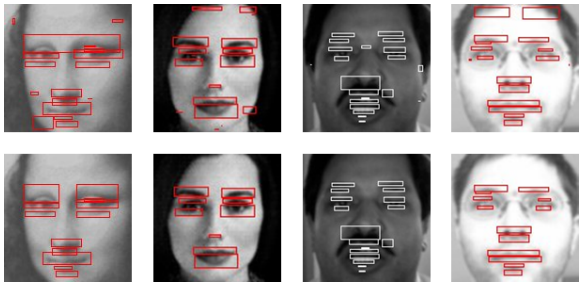


Figure 10: *Les images de la première ligne montrent les boîtes englobantes des CC détectées initialement tandis que celles de la seconde montrent les boîtes des CC conservées.*

trouve au-dessus devient la RAC du nez et l'autre devient la RAC de la bouche. La Figure 11 montre quelques exemples de RAC obtenues en fonction du seuillage de la carte d'énergie.

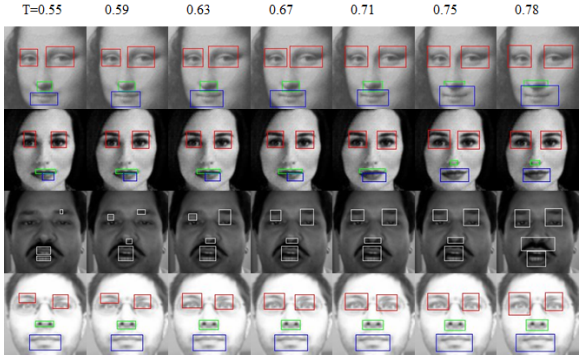


Figure 11: *Détection des RAC en fonction du seuillage de la carte d'énergie.*

3.3. Analyse multi-seuils de la carte d'énergie

Nous avons présenté jusque là comment nous recherchions des régions anatomiques candidates de chacun des yeux, du nez et de la bouche. Toutefois, l'illumination sur un visage donné peut varier, à cause de conditions d'éclairage inégales ou encore d'ombres provenant d'autres objets (cheveux, main...). Un seuillage de la carte d'énergie adéquat pour extraire les yeux n'est pas toujours celui qui permet d'extraire le nez ou la bouche. Ainsi, un seuil doit être choisi pour chaque élément anatomique spécifique du visage. Pour cette tâche, une analyse multi-seuils est proposée. A ce moment précis de notre approche, plusieurs seuils ont été utilisés, chacun permettant d'extraire 4 RAC qui correspondent respectivement aux yeux droit et gauche, au nez et à la bouche. Nous avons généré ainsi 4 ensembles. Le premier contient tous les RAC de l'œil droit, le second, tous les RAC de l'œil gauche et ainsi de suite. Le but de l'analyse multi-seuils est de trouver un seuil adéquat et donc une RAC adéquate pour chaque région spécifique R du visage. Ici, nous supposons que la RAC recherchée soit celle dont la position et la taille varie peu en fonction du seuil t de la carte d'énergie. En effet, si la position et la taille d'une RAC varie peu, cela signifie que l'énergie de la zone définie par la RAC est stable par rapport à t. Ainsi, pour une région donnée R, nous définissons 4 fonctions : $x_R(t)$, fonction des abscisses du point S des RAC, $y_R(t)$, fonction des ordonnées du point S des RAC, $w_R(t)$, fonction des largeurs des RAC et $h_R(t)$, fonction des hauteurs des RAC de la région R. Un seuil adéquat t^* d'une région spécifique R est donné par l'équation (7).

$$\begin{aligned} D(t) &= \left|\frac{\delta}{\delta t}x_R\right| + \left|\frac{\delta}{\delta t}y_R\right| + \left|\frac{\delta}{\delta t}w_R\right| + \left|\frac{\delta}{\delta t}h_R\right| \qquad (7)\\ A(t) &= \beta \cdot W \cdot H - w_R(t) \cdot h_R(t)\\ t^* &= \max_{A(t)>0 \text{ and } D(t)<\varepsilon_R} t \end{aligned}$$

β est le ratio maximum entre l'aire de la région spécifique R et celle de la fenêtre du visage. β dépend de la connaissance que nous avons sur les proportions maximales de chaque région par rapport à l'ensemble du visage. Par exemple, pour un œil, il est de 0,1. Notons que β est une borne supérieure, elle n'est en réalité dans notre méthode que très rarement atteinte, mais nous permet d'exclure quelques valeurs absurdes.

Quant à ε_R, il s'agit de la moyenne des $D(t)$ où les valeurs nulles de $D(t)$ ne sont pas prises en compte. Au final, 4 seuils sont calculés indépendamment et permettent de choisir une RAC pour une région R spécifique comme le montre la Figure 12.

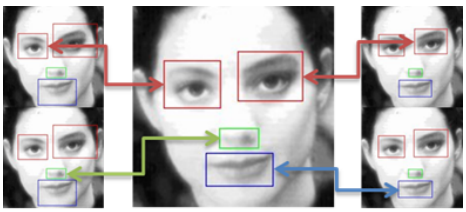


Figure 12: *Sélection des RAC pour chaque région R par l'analyse multi-échelle.*

4. Evaluation

Dans la littérature, rares sont les méthodes qui cherchent à la fois les yeux, le nez et la bouche, ainsi une évaluation générale comparative de l'ensemble est difficile. Par contre, de nombreuses méthodes cherchent à extraire un de ces éléments spécifiques du visage. Pour cette évaluation, nous utilisons les bases BioID et Color FERET. La base d'images BioID est composée de 1520 images. chacune d'entre elles contient un visage. Sur cette base, seuls les yeux, plus précisément l'iris des yeux sont annotés. Les conditions d'illumination ne sont pas contrôlées et varient. De nombreux visages subissent des occlusions (main, lunettes). Cette évaluation se basera essentiellement sur la détection des yeux. En effet, en ce qui concerne la détection des yeux, il est possible de donner non seulement le taux de bonne détection,

mais aussi la précision rattachée à cette détection. Malheureusement, ce n'est ni le cas pour le nez, ni pour la bouche où le seul indicateur sera le taux de bonne détection. Nous méthode est comparée à celles de Li et al. [YLM08] et celle de Asteriadis et al. [SAP09]. Dans [YLM08] et [SAP09], la mesure de Jesorsky [OJF01] est utilisée pour évaluer la précision de la détection des yeux sur la base BioID.

Soient d_r, la distance entre la position réelle du centre de l'iris droit et le centre de la région correspondant à l'œil droit et d_l, l'équivalent pour l'œil gauche. Soit d_{rl} la distance réelles entre les centres de l'iris droit et gauche. Jesorsky definit l'erreur *err* par l'équation (8).

$$err = \frac{max(d_r, d_l)}{d_r l} \tag{8}$$

Pour la tâche qui consiste à détecter les yeux, une erreur de 0,25 est acceptée. En d'autres termes, quand $err < 0,25$, on considère que le détection a réussi. Notons que cet article traite du sujet de la détection des yeux et non celui de la localisation. La différence principale entre détection et localisation est que la première cherche une région alors que la seconde essaie de localiser un point saillant précis. Pour les problèmes de localisation de l'œil, une erreur inférieure à 0,05 ou 0,1 est demandée [XTC09]. Le standard de Jesorsky utilise les positions du centre des iris. Ainsi, nous devons tout d'abord estimer la position de l'iris.

Notre approche qui consiste à détecter la région des yeux n'a pas été exclusivement conçue pour les images de visages frontales. Au fur et à mesure que le visage tourne, seul un des deux yeux devient visible. Notre approche est donc conçue pour détecter au moins un œil. Par contre, les visages dans BioID sont frontales. Ainsi, les deux yeux sont visibles. Nous avons donc calculé le taux d'images dans BioID où un seul œil est détecté. Dans 0,0034% des visages de BioID, notre méthode détecte un œil alors que les deux sont présents et comme le standard de Jesorsky requiert les deux yeux, notre évaluation se portera donc sur 99,9966% des visages.

Dans notre approche, les lignes horizontales sont extraites. Puisque les sourcils ou encore l'arcade sourcilière sont à direction horizontale, ils sont systématiquement inclus dans la région des yeux. Puisque le but est d'extraire les boîtes englobantes des régions anatomiques du visage, nous supposons que le sourcil peut être être pertinent et ainsi peut être incorporé dans la région des yeux. Cependant, comme le centre de l'œil dans BioID est l'iris et comme notre approche inclut les sourcils ou l'arcade sourcilière, nous avons choisi de réduire les RAC des yeux du tiers de leur hauteur. Ainsi, l'abscisse et la largeur des RAC ne changent pas, contrairement à l'ordonnée et la hauteur. Soient y_E et h_E l'ordonnée et la hauteur de la région détectée de l'œil, nous définissons la nouvelle ordonnée y_{eye} et la nouvelle hauteur h_{eye} par l'équation 9.

$$\begin{aligned} y_{eye} &= y_E - \frac{1}{3} \cdot h_E \\ h_{eye} &= \frac{2}{3} \cdot h_E \end{aligned} \tag{9}$$

Alors, les coordonnées du centre *C* de l'œil sont estimées par l'équation 10.

$$\begin{aligned} x_C &= x_E + \frac{1}{2} \cdot w_E \\ y_C &= y_{eye} + \frac{1}{2} \cdot h_{eye} \end{aligned} \tag{10}$$

Méthode	Détection (%)	Erreur moyenne
Li et al.	96	0,1004
Asteridias et al.	96	non indiqué
Notre méthode	97,23	0.1130

Table 1: *Comparaison entre les méthodes de Li et al. Asteriadis et al. et de la nôtre sur BioID.*

Le tableau 1 compare notre méthode avec celles de Li et al. et de Asteriadis et al. Le taux de détection correspond au pourcentage de visages où $err < 0,25$. Notre méthode possède un meilleur taux de détection pour une précision moindre par rapport à Li et al. Malgré l'approximation sur le centre des yeux que nous avons utilisée, les résultats sur BioID sont similaires.

Seuil	Détection (%)	Erreur moyenne
0.39	58.98	0.3872
0.43	62.47	0.3495
0.47	67.95	0.3081
0.47	67.95	0.3081
0.51	72.90	0.2768
0.55	77.37	0.2540
0.59	82.79	0.2127
0.63	88.48	0.1773
0.67	91.67	0.1549
0.71	93.90	0.1417
0.74	**95.39**	**0.1309**
0.78	93.43	0.1434
0.82	87.80	0.1792
0.86	72.22	0.2786
0.90	45.60	0.4971
multi-threshold	**97.23**	**0.1130**

Table 2: *Taux de détection et erreur moyenne relative à la détection des yeux en fonction du seuil appliqué sur la carte d'énergie.*

Notre méthode est basée sur la sélection de candidats adéquats parmi les RAC d'une région saillante du visage. La première question que l'on peut se poser et comment serait la détection sans l'analyse multi-seuils pour évaluer l'intérêt de celle-ci. C'est pourquoi le tableau 2 donne le taux de détection et l'erreur moyenne de la détection des yeux en fonction du seuil utilisé sur la carte d'énergie horizontale.

Comme nous pouvons le voir sur le tableau 2, le pourcentage de bonne détection ($err < 0,25$) n'est pas distribué de manière égale. On observe un maximum du pourcentage de détection et une erreur minimale pour un seuil proche de

0,74. C'est en partie dû à la normalisation de la carte d'énergie horizontale. Cette normalisation a aussi pour effet de réduire l'espace de recherche. Notons que pour un seuil supérieur à 0,86, le taux de détection décroît et l'erreur augmente rapidement. En effet, lorsque le seuil de la carte d'énergie est trop élevé, les CC issues de différents éléments saillants du visage ont tendance à fusionner. Ce phénomène est brutal, puisqu'une toute petite augmentation de ce seuil peut fusionner deux RAC auparavant distincts. Sans l'analyse multi-seuils, pour un seuil fixé à 0,74, la détection de l'œil réussit avec toutefois de bons rappel et précision. Néanmoins, clairement, l'approche multi-seuils montre de meilleurs résultats à la fois en terme de taux de détection, mais aussi de précision que l'utilisation d'un seuil fixe.

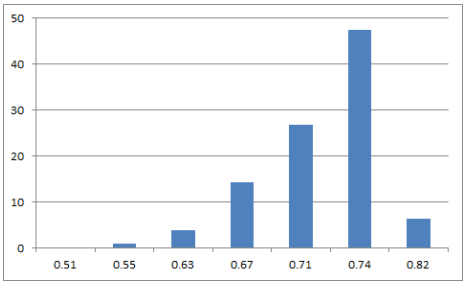


Figure 13: *Pourcentage des seuils choisis par l'analyse multi-seuils.*

La Figure 13 montre le pourcentage des sept seuils choisis par l'analyse multi-seuils pour la détection des yeux. On remarque tout d'abord que le seuil le plus sélectionné est celui qui correspond à la meilleure détection en terme de taux de détection et d'erreur moyenne. Cependant, seul 47% des RAC choisis des yeux sont issues de ce seuil, bien que le taux de bonne détection à ce seuil soit de 95,39%. L'analyse multi-échelle permet de choisir d'autres seuils permettant d'obtenir des RAC de l'œil plus précise pour la détection des yeux. Ceci confirme l'hypothèse que nous avons formulée, à savoir que les RAC adéquats sont celles qui présentent une certaine stabilité de position et de taille malgré la variation du seuil. Enfin, la Figure 13 montre aussi que l'analyse multi-seuils préfère sélectionner les seuils inférieurs à 0,74 plutôt que ceux qui lui sont supérieurs. Comme nous l'avons dit plus tôt, lorsque le seuil atteint une certaine valeur, les CC des RAC sélectionnées ont tendance à fusionner entre elles. Ceci a pour conséquence d'introduire une discontinuité dans la variation de la position et de la taille des éléments détectés.

Nous avons aussi évalué notre méthode sur la base Color FERET. Chaque image contient un visage dans différentes conditions de pose, d'illumination... De nombreuses personnes ont une barbe, une moustache ou encore des lunettes. Les positions de l'iris, du bout du nez, du centre de la bouche sont données sur la majorité des visages frontaux et sur quelques images de visages non frontaux.

Nous avons aussi évalué notre méthode sur la base LFW. Les visages sont au centre de l'image et ont toutes été détectées par la méthode de Viola et Jones. Cette base est particulièrement diversifiée en termes de conditions de prise de vue, d'illumination, de qualité (flou). Il s'agit de l'une des bases les plus difficiles et exigeantes qui existent actuellement. Malheureusement, les positions des différents éléments anatomiques n'y sont pas annotées. La base LFW contient un nombre très important d'images. Il nous était impossible de les annoter toutes. Nous avons choisi d'annoter la position des iris de toutes les images dont le prénom de la personne centrale à l'image commence par un "C" (plus de 900 images).

Base (%)	Détection(%)	erreur moyenne
BioID	97,23	0,1130
Color Feret	97.60	0,1110
LFW	93.74	0,1107

Table 3: *Taux de détection et erreur moyenne sur différentes bases de visages*

Comme le montre le tableau 3, notre méthode parvient à détecter les yeux avec une erreur moyenne semblable pour les trois bases. On remarque que le taux de détection des yeux est plus faible sur la base LFW, ce qui confirme la difficulté de cette base. Les figures 14,15 et 16 montre respectivement des exemples de résultats visuels sur les base BioID, Color Feret et LFW.

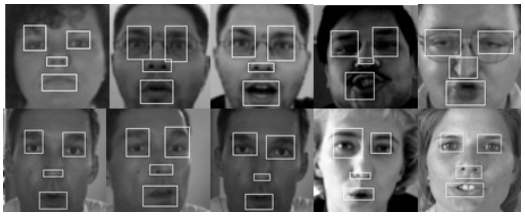


Figure 14: *Résultats visuels sur BioID.*

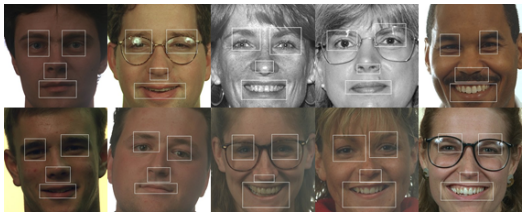


Figure 15: *Résultats visuels sur Color FERET.*

La figure 17 montre des exemples de détections erronées ou incomplètes.

En ce qui concerne la détection du nez et de la bouche, seule le taux de détection a été testé. Notre approche est conçue pour donner au moins une RAC du nez et de la bouche, puisque au moins une partie de ces éléments est supposée visible. Le tableau 4 donne le résultat du taux de détection du nez et de la bouche dans la base Color FERET et MIT/CMU. Ces bases ont l'avantage d'avoir les positions du bout du nez et de la bouche annotées. La détection du nez et celle de la bouche n'ont pas été évaluées sur BioID, car ces éléments ne sont pas annotés dans cette base. Pour ces tests, nous supposons qu'une région est correctement détectée si

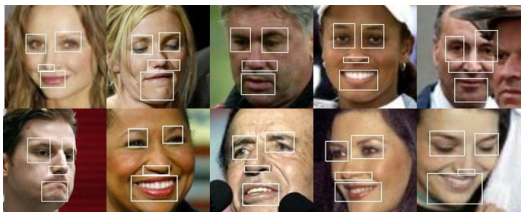


Figure 16: *Résultats visuels sur LFW.*

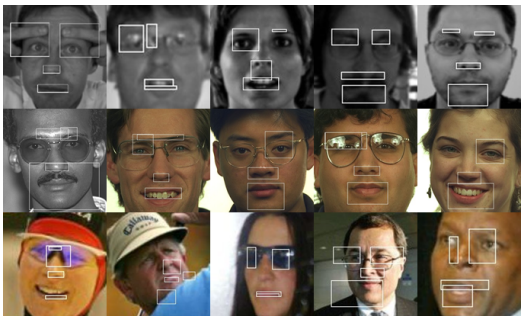


Figure 17: *Exemples de détections erronées ou incomplètes.*

elle contient le point annoté correspondant tout en respectant une contrainte sur l'aire de la région. Pour le nez, l'aire de la région détectée doit être inférieure à 5% de l'aire de la fenêtre du visage et pour la bouche, elle doit être inférieure à 8%. Le tableau 4 montre le taux de détection du nez et de la bouche. Globalement, la bouche semble correctement détectée. Le taux de détection du nez est nettement inférieur, d'une part parce que la détection du nez est difficile lorsque le sujet porte la moustache et d'autre part parce que notre méthode a tendance à détecter la base du nez et non le bout du nez.

Base (%)	Nez(%)	Bouche(%)
Color Feret	75,23	97,50
MIT/CMU	83,63	97,76

Table 4: *Taux de détection du nez et de la bouche dans les bases Color FERET et MIT/CMU.*

Sur la base BioID, Le temps moyen de calcul de la méthode pour un visage est de 16 ms sur un processeur Intel Core i7-2670 QM cadencé à 2,2 GHz. L'extraction des RAC à un seuil donné se fait en moyenne en 2 ms. Ces mesures ont été prises sur un logiciel de test qui n'est pas optimisé. De plus, toute la partie d'extraction des RAC en fonction des seuils qui est la plus longue, car implémentée de manière séquentielle peut être incorporée dans une architecture à processus parallèle. Sinon, le calcul de la carte horizontal d'énergie est relativement rapide, puisque l'image intégrale proposée par Viola et Jones est utilisée. Enfin, la partie concernant l'analyse multi-seuils est la plus rapide dans cette approche. Nous utilisons ici 7 seuils. Pour chaque seuil, 4 régions sont extraites. chaque région génère 4 valeurs (position et taille). Ainsi, l'extraction des 4 seuils adéquats demande un nombre de calculs constants sur un tableau de 112 valeurs.

5. Conclusion

Dans cet article, nous proposons une nouvelle méthode qui utilise un seul type de motif de filtre de Haar à taille adaptative permettant d'extraire les boîtes englobantes des yeux, du nez et de la bouche. Connaissant le niveau d'observation du visage, nous sommes capables de détecter les différentes parties du visage grâce à une carte d'énergie horizontale qui s'avère efficace. Cet article montre aussi comment de simples connaissances sur les visages permet d'améliorer ou de consolider la détection des régions anatomiques faciales. De plus, nous proposons une méthode d'analyse multi-seuils capable de choisir un seuil adéquat pour chaque élément du visage, malgré des conditions d'illumination difficiles. L'évaluation a montré l'efficacité de l'analyse multi-seuil. Elle a aussi montré que la méthode permet de détecter les yeux avec précision, malgré l'approximation relative au sourcil ou à l'arcade sourcilière que nous avons utilisée. La détection de la bouche est aussi bonne tandis que celle du nez reste toujours difficile, en particulier, lorsque le sujet a une moustache et une barbe.

Références

[AB10] AKHLOUFI M., BENDADA A. : Locally adaptive texture features for multispectral face recognition. *Systems, Man and Cybernetics* (octobre 2010), 3308–3314.

[AYH89] A. YUILLE D. C., HALLINAN P. : Feature extraction from faces using deformable templates. *CVPR* (juin 1989), 104–109.

[DPM11] D. PETRISOR C. FOSALAU M. A., MARIUT F. : Algorithm for face and eye detection using colour segmentation and invariant features. *TSP* (2011), 564–569.

[DZZ14] DI ZHU SIYU XIA X. Z., ZHENG J. : Hybrid method for human eye detection. *CCDC* (2014), 5368–5373.

[GS07] GIZATDINOVA Y., SURAKKA V. : Automatic detection of facial landmarks from au-coded expressive facial images. *ICIAP* (septembre 2007), 419–424.

[IC13] INHO CHOI D. K. : Generalized binary pattern for eye detection. *Signal Processing Letters* (2013), 343–346.

[JH09] JIAN W., HONGLIAN Z. : Eye detection based on mult-angle template matching. *Image Analysis and Signal Processing* (avril 2009), 241–244.

[JP09] JAIN A., PARK U. : Facial marks : Soft biometric for face recognition. *ICIP* (novembre 2009), 37–40.

[KP97] KOTROPOULOS C., PITAS I. : Rule-based face detection in frontal views. *Acoustics, Speech and Signal Processing. Vol. 4* (avril 1997), 2537–2540.

[LR12] LAXMI V., RAO P. : Eye detection using gabor filter and svm. *ISDA* (2012), 880–883.

[LZM12] LIN ZHONG QINGSHAN LIU P. Y., METAXAS D. : Leanrning active facial patches for expression analysis. *CVPR* (juin 2012), 2562–2569.

[MCT09] MURPHY-CHUTORIAN E., TRIVEDI M. : Head pose estimation in computer vision : A survey. *Pattern Analysis and Machine Intelligence. Vol. 31*, Num. 14 (avril 2009), 607–626.

[MZW10] MINGCAI ZHOU LIN LIANG J. S., WANG Y. : Aam based face tracking with temporal matching and face segmentation. *CVPR* (novembre 2010), 701–708.

[OJF01] O. JESORSKY K. J. K., FRISHOLZ R. W. : Robust face detection using the hausdorff distance. in : Audio and video based person authentication. *LNCS* (2001), 90–95.

[qZhC12] QING ZHU J., HUI CAI C. : Real-time face detection using gentle adaboost algorithm and nesting cascade structure. *ISPACS* (2012), 33–37.

[SAP09] S. ASTERIADIS N. N., PITAS I. : Facial feature detection using distance vector fields. *Pattern Recognition. Vol. 42*, Num. 7 (2009), 1388–1398.

[TCT01] T. COOTES G. E., TAYLOR C. : Active appearance models. *Pattern Analysis and Machine Intelligence. Vol. 23*, Num. 6 (juin 2001), 681–685.

[TOM02] T. OJALA M. P., MAENPAA T. : Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence* (juillet 2002), 971–987.

[TT10] TAN X., TRIGGS B. : Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Biometrics Compendium. Vol. 19*, Num. 6 (juin 2010), 1635–1650.

[VJ01] VIOLA P., JONES M. : Rapid object detection using a boosted cascade of simple features. *CVPR. Vol. 1* (juin 2001), 511–518.

[VP05] VUKAINOVIC D., PANTIC M. : Fully automatic facial feature point detection using gabor feature based boosted classifiers. *Systems, Man and Cybernetics. Vol. 2* (octobre 2005), 1692–1698.

[XTC09] XIAOYANG TAN FENGYI SONG Z.-H. Z., CHEN S. : Enhanced pictorial structures for precise eye localization under incontrolled conditions. *CVPR* (2009), 1621–1628.

[YLM08] YI LI PENG-FEI ZHAO B.-K. W., MING D. : An improved hybrid projection function for eye precision location. *MIMI. Vol. 4987* (2008), 312–321.

[ZZ12] ZHU S., ZHANG N. : Face detection based on skin color model and geometry features. *ICICEE* (2012), 991–994.

Reconnaissance d'actions humaines 3D par l'analyse de forme des trajectoires de mouvement.

Maxime Devanne[1,2,3], Hazem Wannous[1], Stefano Berretti[3], Pietro Pala[3], Mohamed Daoudi[1,2] et Alberto Del Bimbo[3]

[1]Université de Lille 1 - LIFL (UMR Lille1/CNRS 8022)
[2]Institut Mines-Telecom
[3]Université de Florence, Italie

Résumé

La reconnaissance d'actions humaines dans des séquences vidéo 3D est un problème important, actuellement au cœur de nombreux domaines de recherche comme la vidéo surveillance, les interfaces Homme-Machine et la rééducation. Le développement d'algorithmes de reconnaissance d'actions précis et efficaces est une tâche difficile à cause des fortes variabilités des formes humaines, des vêtements et du mouvement. Dans ce papier, nous proposons un nouvel outil permettant de représenter de manière compacte, de comparer et de reconnaître des actions humaines capturées à partir de caméras de profondeur. Dans un premier temps, les coordonnées 3D de chaque articulation du squelette humain sont considérées comme une chaîne de mouvement. L'évolution spatiale et temporelle de ce vecteur caractéristique est ensuite représentée par une trajectoire dans l'espace des actions. Grâce à cette représentation basée sur les articulations 3D, nous sommes capable de capturer simultanément aussi bien l'apparence géométrique du corps humain que sa dynamique au cours du temps. Le problème de reconnaissance d'actions est ensuite formulé comme un problème de recherche de similarités entre la forme des trajectoires dans une variété riemannienne. La classification par l'algorithme des k-plus-proches-voisins est ensuite effectuée sur la variété pour bénéficier de la géométrie riemannienne dans l'espace des formes. Notre méthode est évaluée sur deux bases de données publiques. En comparaison avec les méthodes existantes dans l'état de l'art, les résultats obtenus montrent l'efficacité de l'approche proposée avec un taux supérieur à 91% *sur les deux bases de données.*

Abstract

Recognizing human actions in a 3D video sequence is an important open problem, which is currently at the heart of many research domains including surveillance, Human-Machine interfaces and rehabilitation.Developing algorithms for action recognition that are both accurate and efficient is challenging due to the variability of the human shape, clothing and motion. In this paper, we propose a new framework which allows compact representation, quick comparison and accurate recognition of human actions in video sequences from depth sensors. Initially, the 3D coordinates of the joints of a human skeleton are considered as one motion channel and the spatial and temporal evolution of this feature vector is represented as a trajectory in an action space. Thanks to such a 3D joint-based framework, we are able to capture both the geometric appearance and the dynamics of the human body simultaneously. The action and activity recognition problem is then formulated as the problem of computing the similarity between the shape of trajectories in a Riemannian manifold. Classification using kNN is finally performed on this manifold taking benefit from Riemannian geometry in the open curve shape space. Experiments on two action datasets, namely MSR Action 3D and UTKinect, are performed. Compared to state-of-the-art methods, results show high performance, above 91%*, on the two challenging datasets.*

Mots clé : Reconnaissance d'actions 3D, modélisation temporelle, espace des formes, variété riemannienne.

1. Introduction

Les technologies de l'imagerie ont récemment montré un avancement conséquent avec l'apparition de nouvelles caméras de profondeur comme la Kinect de Microsoft [Mic13] ou l'Asus Xtion PRO LIVE [ASU13]. Ces nouveaux périphériques d'acquisition ont stimulé le développement de divers applications prometteuses comme l'estimation et la reconstruction de la posture [SFC*11], l'estimation du flot de scène [HB11], la reconnaissance de gestes [RYZ11] et la super-résolution de visages [BDP12]. Une étude récente sur

les applications basées sur les caméras de profondeur peut être trouvée dans [HSXS13]. Les résultats encourageant montrés dans ces travaux peuvent être en partie expliqués par les avantages apportés par de telles caméras de profondeur, comme la segmentation premier-plan/arrière-plan et la robustesse aux changements de conditions d'éclairage. Par conséquent, plusieurs bibliothèques permettant la détection et le suivi du corps humain en temps réel ont vu le jour. Alors que ces méthodes d'extraction et de représentation du corps humain par des silhouettes ou des squelettes ont évolué rapidement, les techniques interprétant la dynamique de ces données, afin de comprendre les actions observées, sont assez limitées. Cette tâche est notamment compliquée par la nécessité d'être invariant aux transformations géométriques ainsi qu'à la vitesse d'exécution de l'action. De plus, d'autres défis importants comme les données bruitées et la variété de poses au sein d'un même type d'action rendent la reconnaissance d'actions d'autant plus difficile.

La reconnaissance d'actions humaines se basant sur l'analyse de données fournies par des caméras de profondeur ont attiré de nombreux groupes de recherche dans les dernières années. Les approches décrites par la suite peuvent être groupées en trois principales catégories selon la méthode d'utilisation de l'information de profondeur : les méthodes basées squelette, les méthodes basées carte de profondeur et les méthodes hybrides qui combinent et exploitent les deux types d'information. Suivant cette catégorisation, les méthodes existantes pour la reconnaissance d'action à partir de caméra de profondeur sont analysées par la suite.

Les approches basées squelette sont devenues populaires suite au travail de Shotton et al. [SFC*11] où une méthode de prédiction précise en temps réel des positions des articulations 3D du corps humain à partir de cartes de profondeur est proposée. S'appuyant sur la position de ces articulations, [XCA12] propose une approche qui calcule des histogrammes des positions de 12 articulations comme une représentation compacte de la posture. Ces histogrammes de posture sont ensuite regroupés en k mots visuels. L'évolution temporelle de ces mots visuels est modélisée par un model de Markov caché. Dans [YT12], la reconnaissance d'actions humaines est obtenue par l'extraction de trois caractéristiques pour chaque articulation basées sur la différence deux à deux des positions des articulations : dans la trame actuelle, entre la trame actuelle et la précédente et entre la trame actuelle et la première de la séquence représentant la posture neutre. L'analyse des composantes principales (PCA) est utilisée pour réduire les redondances et le bruit de ces caractéristiques, et ainsi obtenir une représentation compacte appelée *EigenJoints* pour chaque trame. Finalement, un classifieur bayésien naïf est utilisé pour la classification multi-classes.

Les méthodes basées sur les cartes de profondeur s'appuient sur l'extraction de descripteurs à partir de l'ensemble des points de l'image de profondeur. La modélisation de la dynamique de l'action est ainsi un défi important résolu de manière différente selon les approches. L'approche proposée dans [LZL10] emploie des silhouettes 3D pour décrire les postures et utilise un modèle graphique pour modéliser la dynamique de l'action. Dans [YZT12], la dynamique est décrite par l'intermédiaire de cartes de mouvement de profondeur qui mettent en avant les zones de la scène où un mouvement est effectué. D'autres méthodes proposent de travailler dans un espace 4D divisé en cellules spatiotemporelles pour y extraire des caractéristiques représentant l'apparence de profondeur comme *Spatio-Temporal Occupancy Pattern* [VNO*12], *Random Occupancy Pattern* [WLC*12] and *Depth Cuboid Similarity Feature* [XA13]. Enfin le travail dans [OL13] propose de quantifier l'espace 4D en utilisant les sommets d'un polychore puis en modélisant la distribution des vecteurs normaux pour chaque cellule.

Les solutions hybrides combinent les informations issues des deux flux (squelette et carte de profondeur) pour modéliser l'action. Wang et al. [WLWY12] propose de calculer un descripteur appelé *Local Occupancy Pattern* autour de chaque articulation. Dans [OBT13], les actions sont caractérisées par la combinaison des angles entre articulations calculés sur les squelettes et les histogrammes de gradients orientés calculés sur l'image de profondeur.

Ces approches, basées sur les données de profondeur, bénéficient des nombreux travaux réalisés sur la reconnaissance d'action à partir de vidéos couleur 2D [WRB11, TCSU08, BTR12, Pop10]. Outre les méthodes euclidiennes [SZTL14], des récentes techniques reformulent de manière intéressante le problème de reconnaissance d'actions à travers des espaces non-euclidiens comme les variétés riemanniennes [VRCC05, HSWL12, AAASC11, Lui12], que nous souhaitons exploiter dans ce papier.

2. Vue d'ensemble de l'approche

Une action humaine est naturellement caractérisée par l'évolution du corps humain au cours du temps. Les données de squelettes contenant la position 3D des différentes parties du corps fournissent une représentation précise de la posture du corps humain. Ces caractéristiques peuvent facilement être extraites et suivies à partir des cartes de profondeur. De plus elles fournissent une information locale sur le corps humain. Cependant, même si les positions 3D précises des différentes articulations sont disponibles, la tâche de reconnaissance d'actions reste difficile à cause de variations temporelles et spatiales dans la manière d'effectuer une action.

Ces considérations nous motivent à aborder le problème de la reconnaissance d'action en proposant une approche basée sur l'analyse de l'évolution des articulations du squelette au cours de la séquence vidéo. Pour cela, nous modélisons un squelette par un vecteur multidimensionnel obtenu en concaténant les coordonnées 3D de ses articulations. Ensuite nous considérons la trajectoire que ce vecteur décrit dans l'espace euclidien multidimensionnel modélisant la dynamique de l'ensemble des articulations. Ces trajectoires sont ensuite interprétées dans une variété riemannienne afin de comparer leur forme par l'intermédiaire d'un recalage temporelle dans *l'espace des formes*. De ce fait, nous

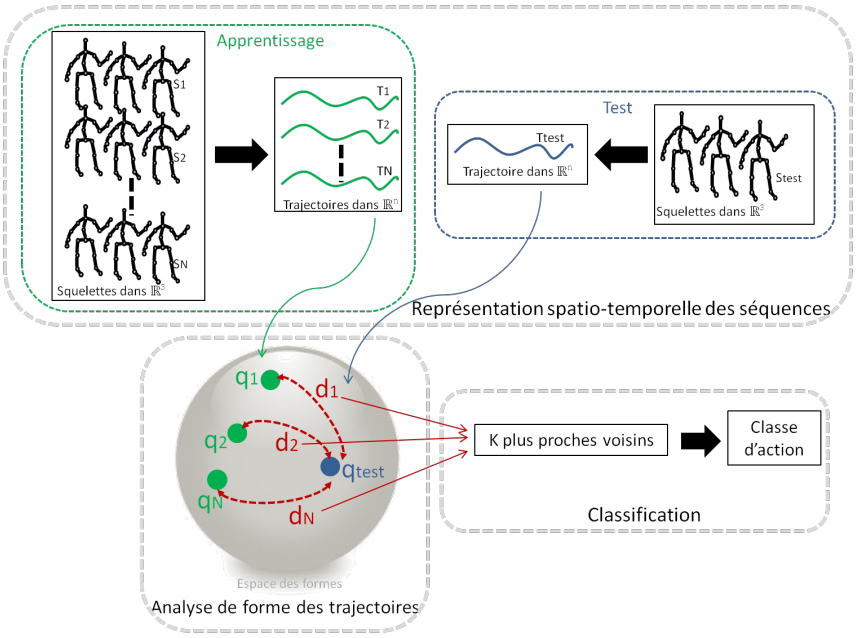


Figure 1: Vue d'ensemble de notre approche. Tout d'abord les séquences de squelettes sont modélisées par des trajectoires dans *l'espace des actions*. Ces trajectoires sont ensuite représentées dans *l'espace des formes*. La reconnaissance est finalement conduite grâce à l'algorithme des *k*-plus-proches-voisins sur cette variété.

reformulons le problème de reconnaissance d'action par une analyse statistique dans *l'espace des formes*. Une métrique élastique est utilisée sur cette variété pour comparer la forme des trajectoires. Cette distance permet à l'approche d'être invariante à l'élasticité des trajectoires. En d'autres termes, elle nous permet de faire face à un défi important de la reconnaissance d'actions : l'invariance à la vitesse d'exécution de l'action. La figure 1 illustre l'approche proposée.

Le reste du papier est organisé comme suit : La section 3 décrit la représentation spatiotemporelle d'une action par une trajectoire. La section 4 introduit l'outil riemannien utilisé pour l'analyse et la comparaison de la forme des trajectoires. Dans la section 5, nous présentons un outil statistique sur la variété riemannienne ainsi que l'algorithme utilisé pour la classification. La section 6 présente les différents résultats obtenus sur deux bases de données publiques. Enfin, la section 7 conclut le papier et discute de futures directions de recherche.

3. Représentation dans l'espace des actions

Grâce aux caméras de profondeur, un squelette humanoïde 3D peut être efficacement extrait à partir des cartes de profondeur depuis l'apparition du travail de Shotton et al. [SFC*11]. Ces squelettes contiennent la position 3D d'un certain nombre d'articulations représentant différentes parties du corps humain. Le nombre d'articulations estimées dépend de l'outil utilisé en combinaison avec le périphérique. Les squelettes extraits grâce au SDK de Microsoft contiennent 20 articulations alors que ceux extraits à partir du SDK de PrimeSense NiTE n'en contiennent que 15. Pour chaque trame t d'une séquence, la position 3D de chaque articulation i est représentée par trois coordonnées exprimées dans le système de référence de la caméra $p_i(t) = (x_i(t), y_i(t), z_i(t))$. Afin de garantir une invariance aux transformations géométriques (rotation et translation), nous alignons l'ensemble des squelettes par rapport à un squelette de référence en calculant la matrice de transformation optimale entre les squelettes. Soit N_j le nombre d'articulations contenues dans un squelette, le vecteur caractéristique à la trame t est défini comme :

$$v(t) = [x_1(t)\ y_1(t)\ z_1(t)\ \ldots\ x_{N_j}(t)\ y_{N_j}(t)\ z_{N_j}(t)]^T . \quad (1)$$

La taille d'un tel vecteur caractéristique est $3N_j$. Pour une séquence de N_f trames, un nombre correspondant de vecteurs colonne est défini et concaténé pour obtenir une matrice caractéristique décrivant la séquence entière :

$$M = \begin{pmatrix} v(1) & v(2) & \ldots & v(N_f) \end{pmatrix} . \quad (2)$$

Cette matrice caractéristique représente l'évolution de la posture au cours du temps, où chaque vecteur colonne v est vu comme un échantillon d'une trajectoire continue dans R^{3N_j}, représentant l'action dans un espace de $3N_j$ dimensions appelé *espace des actions*.

4. Analyse dans l'espace des formes

Une action est une séquence de poses et peut être vue comme le résultat d'un échantillonnage d'une trajectoire continue dans *l'espace des actions*. La trajectoire est définie comme le mouvement au cours du temps des points caractéristiques encodant les coordonnées 3D des articulations du squelette. Soit une trajectoire dans *l'espace des actions* représentée comme une fonction $\beta : I \rightarrow \mathbb{R}^n$, pour $I = [0,1]$. Pour analyser la forme de β, nous représentons la trajectoire

par la *square-root velocity function* (SRVF) $q : I \to \mathbb{R}^n$, définie comme :

$$q(t) \doteq \frac{\dot{\beta}(t)}{\sqrt{\|\beta(t)\|}} , \quad (3)$$

$q(t)$ est une fonction particulière introduite dans [JKSJ07] qui capture la forme de β tout en offrant des facilités de calcul. Comme montré dans [JKSJ07], la norme $\mathbb{L}^2$ représente la métrique pour comparer la forme de deux trajectoires. L'ensemble de toutes les trajectoires, noté $\mathcal{C}$, est ainsi défini comme :

$$\mathcal{C} = \{q : I \to \mathbb{R}^n | \ \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^n) , \quad (4)$$

Avec la norme $\mathbb{L}^2$ sur son plan tangent, cette variété $\mathcal{C}$ devient alors une variété riemannienne. Chaque élément de $\mathcal{C}$ représente une trajectoire. On définit la distance entre deux éléments q_1 et q_2 par la longueur du chemin géodésique entre q_1 et q_2 sur la variété $\mathcal{C}$. Comme ces éléments ont une norme $\mathbb{L}^2$ unitaire, $\mathcal{C}$ peut être vu comme une hypersphère de l'espace de Hilbert. Ainsi la distance géodésique entre q_1 et q_2 est définie comme :

$$d_c(q_1, q_2) = cos^{-1}(\langle q_1, q_2 \rangle) . \quad (5)$$

Afin de garantir l'invariance à la vitesse d'exécution de l'action, nous devons comparer la forme des trajectoires indépendamment de leur élasticité. Cela nécessite une invariance à la re-paramétrisation des trajectoires. Nous définissons la classe d'équivalence de la forme q par :

$$[q] = \{\sqrt{\dot{\gamma}(t)}(q \circ \gamma)(t) | \ \gamma \in \Gamma\} . \quad (6)$$

où Γ est le groupe de re-paramétrisation. Tous les éléments de cette classe d'équivalence, auxquels est associée une certaine re-paramétrisation γ, sont équivalents à la forme de q. L'ensemble de telles classes d'équivalences est appelé *espace des formes* et noté $\mathcal{S}$. La comparaison entre deux éléments q_1 et q_2 requiert une comparaison entre leur classe d'équivalence $[q_1]$ et $[q_2]$. La distance géodésique dans $\mathcal{S}$ devient ainsi :

$$d_s([q_1], [q_2]) = d_c(q_1, q_2^*) . \quad (7)$$

où q_2^* est l'élément q_2 re-paramétré par rapport à q_1. En pratique, la programmation dynamique est utilisée pour trouver la re-paramétrisation optimal entre deux éléments.

5. Reconnaissance d'action dans l'espace des formes

La méthode proposée pour la reconnaissance d'actions est basée sur l'algorithme des k-plus-proches-voisins appliqué dans *l'espace des formes* sur l'ensemble des données d'apprentissage ou sur des séquences représentatives calculées à l'aide de la moyenne de Karcher [Kar77].

5.1. Calcul de trajectoires moyennes

Un des avantages de l'utilisation d'une approche riemannienne pour la reconnaissance d'action est que cela nous permet d'exploiter des outils statistiques sur les éléments de la variété. Par exemple, nous pouvons utiliser la notion de moyenne de Karcher [Kar77] pour calculer des trajectoires moyennes à partir d'un ensemble de trajectoires. Ainsi, une trajectoire moyenne peut être d'une part calculée à partir d'un ensemble de trajectoires différentes pour représenter une trajectoire intermédiaire. D'autre part, elle peut être calculée à partir d'un ensemble de trajectoires similaires afin d'obtenir un modèle moyen qui peut être vue comme une trajectoire représentative de l'ensemble. De plus, pour classifier une séquence, la distance géodésique doit être calculée avec toutes les séquences d'apprentissage. Pour un grand nombre de séquences d'apprentissage, cela implique un temps de calcul élevé. Utiliser des trajectoires moyennes représentatives peut ainsi diminuer le nombre de séquences d'apprentissage et donc le temps de calcul. Pour un ensemble de trajectoires d'apprentissage représentées dans *l'espace des formes* $q_1, \ldots, q_n$, leur moyenne de Karcher peut être définie comme :

$$\mu = \arg\min \sum_{i=1}^{n} d_s([q], [q_i])^2 . \quad (8)$$

La figure 2 présente un exemple de calcul de moyenne de Karcher pour cinq trajectoires ($q_1 \ldots q_5$). Dans l'étape initiale, q_1 est sélectionné comme la trajectoire moyenne. De manière itérative, la moyenne est mise à jour grâce à la métrique élastique calculée entre toutes les trajectoires q_i. Après convergence, la trajectoire moyenne est donné par q_m.

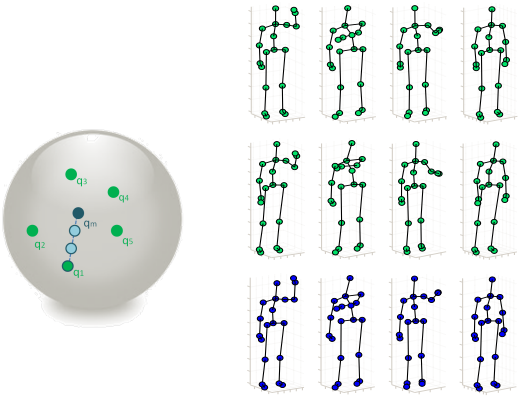


Figure 2: Calcul de la moyenne de Karcher entre 5 trajectoires représentées dans *l'espace des formes* (à gauche) et dans *l'espace des actions* (à droite). Les deux premières séquences correspondent aux trajectoires q_1 et q_2 tandis que la séquence du bas correspond à la trajectoire moyenne q_m.

En calculant une telle trajectoire moyenne pour chaque classe, nous supposons qu'il existe uniquement un seul moyen d'effectuer chaque action, ce qui n'est pas forcément vrai dans tous les cas. Par exemple, une personne gauchère et une personne droitière effectueront différemment une même action. Dans ce cas, calculer une unique trajectoire moyenne pour cette action peut donner une trajectoire non représentative de l'action. Pour cette raison, nous calculons des trajectoires moyennes au sein d'une même classe pour chaque sujet séparément. Au lieu d'avoir une seule trajectoire modèle par classe, nous avons maintenant un modèle par action et par sujet. Ainsi, les différents moyens d'effectuer une même

action sont maintenant séparés et les trajectoires moyennes résultantes sont des trajectoires représentatives des actions.

5.2. k-plus-proches-voisins avec la distance élastique

Soit $\{(X_i,y_i)\}$, $i = 1,\ldots,N$, un ensemble d'apprentissage où X_i appartient à *l'espace des formes* $\mathcal{S}$, et y_i est le label de classe prenant une valeur de $\{1,\ldots,N_c\}$, avec N_c le nombre de classes. L'objectif est de trouver une function $F(X) : \mathcal{S} \longmapsto \{1,\ldots,N_c\}$ pour regrouper les données représentées dans *l'espace des formes* à partir des données labélisées de l'ensemble d'apprentissage. Pour cela, nous proposons d'utiliser le classifieur des k-plus-proches-voisins sur la variété riemannienne appris sur des séquences représentées dans *l'espace des formes*. Cette méthode d'apprentissage nous permet d'exploiter les propriétés géométriques de *l'espace des formes*, et plus particulièrement sa métrique élastique. La classification repose sur le calcul des distances géodésiques entre une trajectoire de test et l'ensemble des trajectoires d'apprentissage. Plus précisément, un ensemble de trajectoires d'apprentissage $X_i : i = 1,\ldots,N$ est représenté comme un ensemble d'éléments $q_i : i = 1,\ldots,N$ dans *l'espace des formes*. Ensuite, pour une séquence de test représentée dans *l'espace des formes*, les distances géodésiques avec l'ensemble des séquences d'apprentissage sont calculées pour trouver les k séquences les plus similaires. Le label de classe associé à la séquence de test est celui le plus représenté parmis les k plus proches séquences d'apprentissage.

6. Résultats expérimentaux

La performance de notre approche est évaluée et comparée avec les méthodes existantes de l'état de l'art sur deux bases de données publiques : MSR Action 3D [LZL10] et UTKinect [XCA12].

6.1. MSR Action 3D

Cette base de donnée publique a été collectée par Microsoft Research [LZL10] et peut être vue comme une base référence pour la reconnaissance d'actions, utilisée dans de nombreux travaux. Elle inclut 20 actions effectuées par 10 personnes 2 ou 3 fois. Au total, 567 séquences sont disponibles. Les différentes actions orientées jeux vidéo sont choisies pour couvrir différentes variations du mouvement des bras, des jambes, du torse et de leur combinaison. Chaque sujet est positionné au centre de la scène face à la caméra. Il était demandé aux personnes d'effectuer les actions avec le bras droit ou la jambe droite lorsque l'action nécessite un seul membre. De plus, toutes les actions sont effectuées sans interactions avec des objets. Les deux principaux défis de cette base de données sont la forte similarité entre certaines actions et la variation de vitesse d'exécution de l'action. Pour chaque séquence les informations de couleur, de profondeur et de squelette sont fournies. Dans notre cas, seules les données de squelette sont utilisées. Comme reporté dans [WLWY12], 10 actions ne sont pas utilisées dans les tests car les squelettes sont soit manquants, soit trop bruités. Pour nos expérimentations, nous utilisons donc 557 séquences.
Nous testons notre approche avec ses différentes méthodes mentionnées dans la section 5.1. Les résultats sont reportés dans la table 1.

Table 1: MSR Action 3D. Nous testons notre approche avec ses différentes méthodes de classification ($k$pp, $k$pp et moyenne de Karcher par action, $k$pp et moyenne de Karcher par action et par sujet.).

Methode	Taux (%)
kpp	88.3
kpp & moyenne de Karcher par action	89.0
kpp & moyenne de Karcher par action/subject	**92.1**

En analysant les résultats, nous pouvons remarquer que le meilleur taux de reconnaissance est obtenu en utilisant la notion de moyenne de Karcher par action et par sujet. En comparaison avec l'utilisation de la moyenne de Karcher par action uniquement, on peut voir que séparer les différentes façons d'effectuer une même action permet d'augmenter le taux de reconnaissance. De plus, les résultats montrent que l'utilisation de trajectoires moyennes est plus efficace que d'utiliser l'ensemble des séquences d'apprentissage. Cela peut s'expliquer pour le cas d'actions similaires. Dans ce cas, une séquence appartenant à une première classe peut être très proche de séquences appartenant à une seconde classe, et ainsi sélectionnée comme un faux positif lors de la classification. Le calcul de trajectoires moyennes peut ainsi augmenter la distance inter-classes et donc améliorer le taux de classification. Par exemple, les deux premières actions de la base (*high arm wave* and *horizontal high arm wave*) sont très proches. Utiliser de telles trajectoires moyennes permet de réduire la confusion entre ces deux actions. Ceci peut être visualiser dans la figure 3 représentant la matrice de confusion pour les deux méthodes de classification différentes.

Dans un deuxième temps, nous comparons notre approche aux autres méthodes de l'état de l'art. Les résultats sont reportés dans la table 2. Pour une comparaison équitable, nous utilisons le même protocole expérimental que les travaux évalués sur la base MSR Action 3D. Les cinq premiers sujets sont choisis pour l'apprentissage, les cinq autres pour le test. En analysant les résultats, nous pouvons voir que notre méthode dépasse les méthodes de l'état de l'art exceptée celle proposée dans [OBT13]. Cependant, cette aproche utilise une combinaison des informations de squelette et de profondeur. Ils reportent qu'en utilisant uniquement les données de squelette, leur méthode donne un taux de reconnaissance de 83.5% plus bas que le taux obtenu avec notre approche.

Pas la suite, nous conduisons les mêmes expérimentations avec toutes les combinaisons possibles de choisir la moitié des sujets comme apprentissage et l'autre moitié comme test. Pour chacune des 252 combinaisons, nous utilisons la moyenne de Karcher par action et par sujet pour l'ensemble d'apprentissage. Nous obtenons un taux de reconnaissance moyen de $87.28 \pm 2.41\%$ (moyenne $\pm$ écart-type). Parmi les 252 combinaisons, le plus bas taux de reconnaissance obtenu est de 81.31% alors que le plus élevé est de 93.04%. En comparaison avec le travail présenté dans [OL13], où le taux moyen est aussi calculé pour toutes les combinaisons possibles, nous dépassons leur résultat de $82.15 \pm 4.18\%$.

Table 2: MSR Action 3D. Comparaison de notre méthode avec les méthodes les plus pertinentes de l'état de l'art.

Methode	Taux (%)
EigenJoints [YT12]	82.3
STOP [VNO*12]	84.8
DMM & HOG [YZT12]	85.5
Random Occupancy Pattern [WLC*12]	86.5
Actionlet [WLWY12]	88.2
DCSF [XA13]	89.3
JAS & HOG^2 [OBT13]	**94.8**
HON4D [OL13]	88.9
Ours	**92.1**

De plus, la faible valeur de l'écart-type dans nos expérimentations montre que notre méthode est très peu dépendante des données choisies pour l'apprentissage. Afin de présenter le taux de reconnaissance obtenu par notre méthode pour chaque action séparément, les matrices de confusion sont calculées pour chacun des cas et présentées dans la figure 3.

Nous pouvons remarquer qu'un très faible taux de reconnaissance est obtenu pour les actions *hammer* et *hand catch*. Cela peut être expliqué par le fait que ces actions sont similaires à d'autres actions. De plus, la façon d'exécuter ces actions varie considérablement selon les sujets. Par exemple, pour l'action *hammer*, les sujets de l'ensemble d'apprentissage ne donnent qu'un seul coup de marteau alors que certains sujets de l'ensemble de test en donnent plusieurs. Dans ce cas, les formes des trajectoires sont très différentes et les séquences correspondantes ne sont pas détectées comme similaires. La figure 4 illustre cet exemple. Comme il est difficile de visualiser les trajectoires dans un espace de grande dimension, ces dernières sont ici représentées en trois dimensions correspondant à une seule articulation (main droite). Les quatre trajectoires correspondent à des échantillons différents de l'action *hammer* où un seul coup de marteau est donné pour les deux premiers cas alors que plusieurs coups sont donnés pour les deux derniers cas. Nous pouvons remarquer que la forme des trajectoires est ainsi différente.

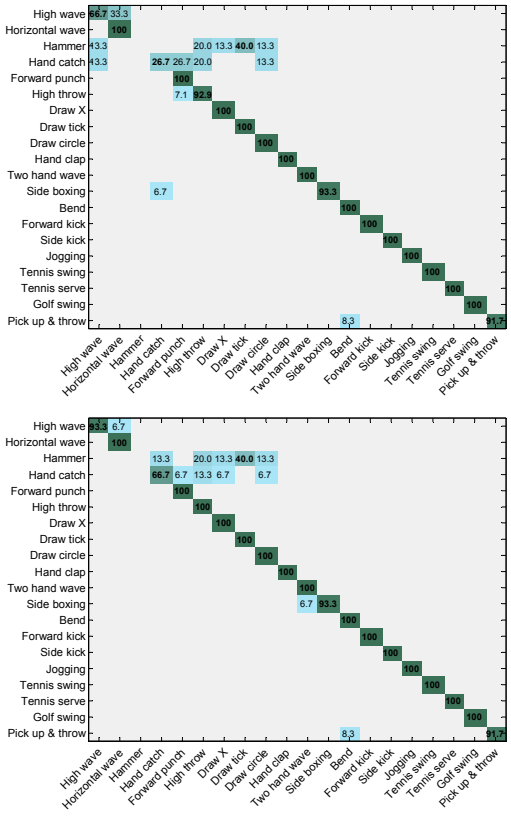


Figure 3: MSR Action 3D. Matrices de confusion obtenues avec notre approche. La classification est effectuée à l'aide le l'algorithme des k-plus-proches-voisins apris sur l'ensemble des séquences d'apprentissage (en haut) et sur des trajectoires moyennes par action et par sujet calculées grâce à la moyenen de Karcher (en bas).

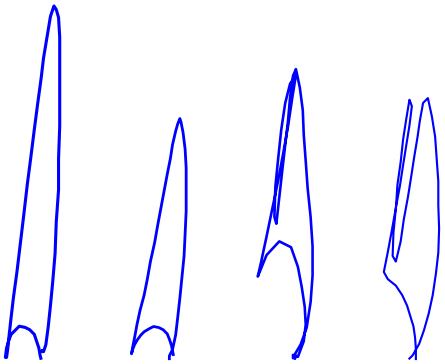


Figure 4: Visualisation d'un cas d'échec pour l'action *hammer*. Quatre trajectoires de la main droite sont représentées. Un seul coup de marteau est donnée pour les deux premières trajectoires tandis que deux coups sont donnés pour les deux trajectoires de droite.

6.2. UTKinect

Afin de confirmer l'efficacité de notre approche, nous proposons également de l'évaluer sur une seconde base de données appelée UTKinect [XCA12] qui présente d'autres défis. Dans cette base, 10 sujets effectuent 10 actions différentes deux fois pour un total de 200 séquences. La base présente trois défis principaux : tout d'abord, les actions sont capturées de différents points de vue ; ensuite, certaines actions nécessitent une interaction avec des objets ; enfin, une autre difficulté est ajoutée avec la présence d'occultations causées par les objets de la scène ou par le champ de vision restreint.

Pour être comparable avec le travail dans [XCA12], nous suivons le même protocole expérimental (leave-one-out-cross-validation). A chaque itération, une séquence est utilisée comme test et toutes les autres pour l'apprentissage. L'opération est répétée afin que chaque séquence soit utilisée une fois comme test. Nous obtenons un taux de reconnaissance de 91.5%, plus élevé que le taux reporté dans [XCA12] s'élevant à 90.9% . Cela montre que notre méthode est robuste aux changements de points de vue et aux occultations de certaines parties du corps. Cependant, en

analysant la matrice de confusion présentée dans la figure 5, nous pouvons remarquer que les plus faibles taux sont obtenus pour les actions utilisant des objets comme *carry* et *throw*. Ces actions sont confondues avec des actions similaires mais sans objet comme *walk* et *push*, respectivement. Cette limite est due au fait que notre approche ne prend en compte que les données de squelette. Ainsi, aucune information à propos de l'objet tenu par le sujet n'est disponible.

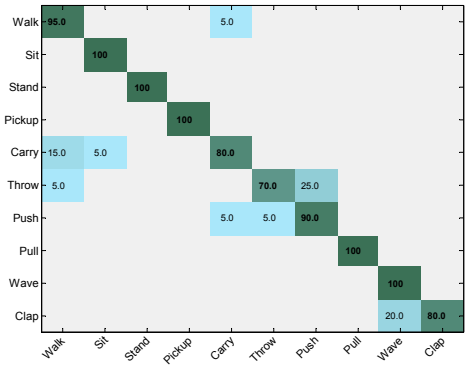


Figure 5: UTKinect. Matrice de confusion obtenue avec notre approche

7. Conclusion

Une approche efficace de reconnaissance d'actions humaines est proposée en utilisant une modélisation spatiotemporelle de trajectoires de mouvement dans une variété riemannienne. La position 3D de chaque articulation du squelette à chaque trame de la séquence est concaténée pour ainsi représenter l'action comme une trajectoire de mouvement dans *l'espace des actions*. Chaque trajectoire est ensuite exprimée comme un élément dans une variété riemannienne appelée *espace des formes*. Grâce à la géométrie riemannienne de la variété, la classification de l'action est résolue grâce l'algorithme des *k*-plus-proches-voisins en utilisant la distance élastique entre les formes des trajectoires. Les résultats expérimentaux sur deux bases de données MSR Action 3D et UTKinect démontrent que notre méthode dépasse les méthodes existantes de l'état de l'art dans la plupart des cas. Comme perspectives, nous envisageons tout d'abord d'intégrer dans notre approche d'autres descripteurs basés sur l'image de profondeur afin de gérer les cas d'interaction avec des objets. De plus, nous souhaitons approfondir les cas d'échec comme les répétitions de gestes pour fournir une approche plus robuste à ces variations. Enfin, nous réfléchissons à de possibles applications de notre travail notamment dans le domaine de la thérapie physique et de la rééducation assistée.

Références

[AAASC11] ABDELKADER M. F., ABD-ALMAGEED W., SRIVASTAVA A., CHELLAPPA R. : Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. *Computer Vision and Image Understanding. Vol. 115*, Num. 3 (2011), 439–455.

[ASU13] ASUS XTION PRO LIVE, : http ://www.asus.com/multimedia/xtion_pro/, 2013.

[BDP12] BERRETTI S., DEL BIMBO A., PALA P. : Superfaces : A super-resolution model for 3D faces. In *Proc. Work. on Non-Rigid Shape Analysis and Deformable Image Alignment* (Florence, Italy, Oct. 2012), pp. 73–82.

[BTR12] BIAN W., TAO D., RUI Y. : Cross-domain human action recognition. *IEEE Trans. on Systems, Man, and Cybernetics, Part B : Cybernetics. Vol. 42*, Num. 2 (avril 2012), 298–307.

[HB11] HADFIELD S., BOWDEN R. : Kinecting the dots : Particle based scene flow from depth sensors. In *Proc. Int. Conf. on Computer Vision* (Barcelona, Spain, Nov. 2011), pp. 2290–2295.

[HSWL12] HARANDI M. T., SANDERSON C., WILIEM A., LOVELL B. C. : Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures. In *Proc. IEEE Work. on the Applications of Computer Vision* (Washington, DC, USA, 2012), WACV'12, IEEE Computer Society, pp. 433–439.

[HSXS13] HAN J., SHAO L., XU D., SHOTTON J. : Enhanced computer vision with microsoft kinect sensor : A review. *IEEE Trans. on Cybernetics. Vol. 43*, Num. 5 (2013), 1318–1334.

[JKSJ07] JOSHI S. H., KLASSEN E., SRIVASTAVA A., JERMYN I. : A novel representation for riemannian analysis of elastic curves in R^n. In *Proc IEEE Int. Conf. on Computer Vision and Pattern Recognition* (Minneapolis, MN, USA, June 2007), pp. 1–7.

[Kar77] KARCHER H. : Riemannian center of mass and mollifier smoothing. *Comm. on Pure and Applied Math.. Vol. 30* (1977), 509–541.

[Lui12] LUI Y. M. : Tangent bundles on special manifolds for action recognition. In *IEEE Trans. on Circuits and Systems for Video Technology* (2012), vol. 22, pp. 930–942.

[LZL10] LI W., ZHANG Z., LIU Z. : Action recognition based on a bag of 3D points. In *Proc. Work. on Human Communicative Behavior Analysis* (San Francisco, California, USA, June 2010), pp. 9–14.

[Mic13] MICROSOFT KINECT : http ://www.microsoft.com/en-us/kinectforwindows/, 2013.

[OBT13] OHN-BAR E., TRIVEDI M. M. : Joint angles similarities and HOG^2 for action recognition. In *Proc. CVPR Work. on Human Activity Understanding from 3D Data* (Portland, Oregon, USA, June 2013), pp. 465–470.

[OL13] OREIFEJ O., LIU Z. : HON4D : Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proc. Int. Conf. on Computer Vision and*

Pattern Recognition (Portland, Oregon, USA, June 2013), pp. 716–723.

[Pop10] POPPE R. : A survey on vision-based human action recognition. *Image Vision Comput.. Vol. 28*, Num. 6 (juin 2010), 976–990.

[RYZ11] REN Z., YUAN J., ZHANG Z. : Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proc. ACM Int. Conf. on Multimedia* (Scottsdale, Arizona, USA, Nov. 2011), pp. 1093–1096.

[SFC*11] SHOTTON J., FITZGIBBON A., COOK M., SHARP T., FINOCCHIO M., MOORE R., KIPMAN A., BLAKE A. : Real-time human pose recognition in parts from single depth images. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition* (Colorado Springs, Colorado, USA, June 2011), pp. 1–8.

[SZTL14] SHAO L., ZHEN X., TAO D., LI X. : Spatio-temporal laplacian pyramid coding for action recognition. *IEEE Trans. on Cybernetics. Vol. PP*, Num. 99 (2014), 1–1.

[TCSU08] TURAGA P., CHELLAPPA R., SUBRAHMANIAN V. S., UDREA O. : Machine recognition of human activities : A survey. *IEEE Trans. on Circuits and Systems for Video Technology. Vol. 18*, Num. 11 (novembre 2008), 1473–1488.

[VNO*12] VIEIRA A. W., NASCIMENTO E. R., OLIVEIRA G. L., LIU Z., CAMPOS M. F. : STOP : Space-time occupancy patterns for 3D action recognition from depth map sequences. In *Iberoamerican Congress on Pattern Recognition* (Buenos Airies, Argentina, Sept. 2012), pp. 252–259.

[VRCC05] VEERARAGHAVAN A., ROY-CHOWDHURY A., CHELLAPPA R. : Matching shape sequences in video with applications in human movement analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 27*, Num. 12 (2005), 1896–1909.

[WLC*12] WANG J., LIU Z., CHOROWSKI J., CHEN Z., WU Y. : Robust 3D action recognition with random occupancy patterns. In *Proc. Europ. Conf. on Computer Vision* (Florence, Italy, Oct. 2012), pp. 1–8.

[WLWY12] WANG J., LIU Z., WU Y., YUAN J. : Mining actionlet ensemble for action recognition with depth cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (Providence, Rhode Island, USA, June 2012), pp. 1–8.

[WRB11] WEINLAND D., RONFARD R., BOYER E. : A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding. Vol. 115*, Num. 2 (février 2011), 224–241.

[XA13] XIA L., AGGARWAL J. K. : Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proc. CVPR Work. on Human Activity Understanding from 3D Data* (Portland, Oregon, USA, June 2013), pp. 2834–2841.

[XCA12] XIA L., CHEN C.-C., AGGARWAL J. K. : View invariant human action recognition using histograms of 3D joints. In *Proc. Work. on Human Activity Understanding from 3D Data* (Providence, Rhode Island, USA, June 2012), pp. 20–27.

[YT12] YANG X., TIAN Y. : Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Proc. Work. on Human Activity Understanding from 3D Data* (Providence, Rhode Island, June 2012), pp. 14–19.

[YZT12] YANG X., ZHANG C., TIAN Y. : Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proc. ACM Int. Conf. on Multimedia* (Nara, Japan, Oct. 2012), pp. 1057–1060.

Un système de suivi multi-objets utilisant une stratégie d'association en trois passes adapté à la vidéosurveillance

M.Rogez[1,2,3] L. Robinault[1,3] et L.Tougne[1,2]

[1]Université de Lyon, CNRS
[2]Université Lyon 2, LIRIS, UMR5205, F-69676, France
[3]Foxstream, Vaulx-en-Velin, France

Résumé

*Le suivi multi-objets est une des thématiques centrales de l'analyse vidéo du fait de son large champ d'application. Nous nous intéressons ici plus particulièrement aux applications en vidéo-surveillance. Ainsi, nous décrivons un ensemble d'améliorations destinées à l'algorithme de suivi multi-objets proposé par [LFP*13]. En particulier, nous généralisons le suivi en retirant la spécialisation faite pour les piétons ; nous intégrons le modèle de scène et de visualisation développé dans [RTR13] afin de permettre un raisonnement tridimensionnel permettant de mieux gérer les occultations ; et enfin nous améliorons le mécanisme de formation et destruction des groupes d'objets grâce à l'introduction d'une passe d'association supplémentaire ainsi que d'un critère de similarité de recouvrement. Enfin, nous évaluons le système proposé sur des vidéos synthétiques et réelles afin de montrer l'apport de nos modifications. L'algorithme proposé améliore sensiblement les performances générales par rapport à la version originale, notamment pour la création et destruction des groupes, et ouvre la possibilité d'un raisonnement tridimensionnel.*

Mots clé : Suivi multi-objets, vidéo-surveillance, groupes, automate fini

1. Introduction

Ce document traite du suivi multi-objets en ligne à partir d'une caméra unique, fixe et calibrée. Plus précisément, nous destinons ces travaux aux applications en vidéosurveillance, où s'appliquent des contraintes de traitement en temps réel (et potentiellement de plusieurs flux sur un même ordinateur), ainsi que la nécessité de pouvoir suivre simultanément différents types d'objets (piétons, voitures ou camions par exemple). Ce sujet est très important en analyse vidéo, car il permet d'obtenir les trajectoires des objets suivis et ainsi autorise une prise de décision basée sur des informations de plus haut niveau que la simple image brute (par exemple pour la détection de trajectoire anormale, ou celle des points d'entrée et de sortie de la scène). En pratique, cependant, un certain nombre de difficultés rendent l'obtention de ces trajectoires difficile : par exemple, comment détecter les objets à suivre ? Comment être robuste face aux mauvaises détections ? Comment gérer les occultations inter-objets ? Comment garantir le maintien de l'identité de chaque objet même après occultation ?

La plupart des algorithmes récents de suivi multi-objets se spécialisent dans le suivi de piétons [BRL*09, FJ14] et ainsi utilisent des détecteurs spécialisés à cet effet, en général en utilisant des HOG et une classification à base de SVM. Cette spécialisation permet d'être robuste face aux fausses détections, aux changements de luminosité et aux occultations partielles. En revanche, ces approches sont souvent coûteuses en temps de calcul, et nécessitent un apprentissage souvent délicat en pratique étant donné la variabilité des points de vue et des poses à considérer. En particulier, cette spécificité de détection peut être vue comme une faiblesse dans certains contextes, comme la détection d'intrusion, où les individus peuvent pénétrer en rampant par exemple, ce qui peut mettre en défaut les détecteurs classiques, car entrainés sur des exemples de piétons debouts. Ou bien, lorsque justement on souhaite réaliser un suivi générique, multi-classes (piétons, voitures et camions par exemple).

Une alternative aux détecteurs spécialisés, mentionnés ci-dessus, est d'utiliser un modèle de fond afin de détecter les objets d'intérêts. Cette approche est largement utilisée en pratique [PA10, CFPV10, RKDL12, GMG12, LFSV12, LFP*13], car elle est souvent plus rapide et plus générique que les détecteurs spécialisés. Cependant, l'utilisation d'un modèle de fond génère souvent plus d'erreurs liées aux changements des conditions d'éclairage, ou aux mouvements de la caméra ou du fond. De plus, raisonner sur un masque de détection pose un certain nombre de contraintes nouvelles par rapport à l'utilisation d'un détecteur spécialisé (voir figure 1) : les composantes connexes extraites du masque de détection (que l'on désignera par "Blob" dans ce qui

suit) peuvent ne pas représenter un seul objet réel, mais un groupe d'objets (blob 1 dans la figure 1). Les causes peuvent être multiples : occultation totale, occultation partielle, forte proximité. De plus, il est possible qu'un objet réel ne soit pas représenté par un seul blob, mais morcelé en plusieurs petits blobs (blobs 3, 4 et 5 dans la figure 1). C'est pourquoi, les approches utilisant un masque de détection doivent résoudre explicitement ces problèmes.

Les travaux présentés dans [LFP*13], sont ceux sur lesquels nous nous appuyons, car ils proposent une formulation adaptée au suivi d'objets à partir de détections sur un masque de mouvement. En particulier, ils proposent un algorithme d'associations objets suivis/blobs détectés capable de gérer les associations multiples qui peuvent se manifester lors d'occultations ou à cause de défauts de segmentation. De plus, ils utilisent la notion de groupe afin de suivre collectivement les objets occultants et occultés pendant la durée de l'occultation ; les identités des individus formant le groupe sont cependant conservées afin de pouvoir les réassigner correctement à la fin de l'occultation. Enfin, ils modélisent l'évolution des objets suivis à l'aide d'un automate fini. Cette modélisation a plusieurs intérêts : d'une part, c'est une représentation compacte et directement intelligible de l'état d'un objet (à l'inverse d'une densité de probabilité par exemple). D'autre part, cette modélisation permet d'adapter les traitements et les valeurs de paramètres spécifiquement pour chaque objet en fonction de son état.

Bien que largement inspirée par les travaux de [LFP*13], dont la formulation est particulièrement adapté à la vidéosurveillance, notre méthode se distingue par les changements majeurs qui suivent :

- **Généralisation du suivi :** nous retirons la spécialisation du suivi aux seuls piétons. Ceci inclut notamment la suppression de la classification des objets et blobs, et donc l'apprentissage associé.
- **Prise en compte de la 3d :** nous tirons parti de la calibration de la caméra pour opérer en 3d, notamment en estimant la position au sol, ainsi que les dimensions réelles des objets suivis.
- **Prise en compte des bâtiments environnants :** en utilisant les travaux de [RTR13], nous avons intégré la connaissance des bâtiments environnants afin de délimiter la zone effective de suivi.
- **Amélioration de la formation/destruction des groupes :** d'une part, un nouvel algorithme d'association incluant une première passe permettant de réaliser les associations très probables a été proposé. Cette passe supplémentaire permet d'évacuer les associations simples très probables qui pourraient autrement interférer avec les formations/destructions de groupe. D'autre part, nous avons introduit, un critère de similarité supplémentaire, basé sur le recouvrement entre la silhouette de l'objet et le blob considéré. Ce critère additionnel est utilisé spécifiquement pour détecter les associations multiples, à la place du critère de forme original peu adapté dans ce cas.
- **Possibilité de re-rentrer dans la scène :** dans la version originale, l'automate fini n'autorise pas un objet sortant à re-rentrer dans la scène, alors qu'en pratique cette transition s'avère utile.

Le reste du document s'organise de la manière suivante : un aperçu global de l'algorithme est présenté dans la partie 2.1. Nous présentons ensuite les différents composants de cet algorithme dans les sections suivantes : à savoir, l'évaluateur de similarité en section 2.2, l'algorithme d'association en section 2.3, la procédure de mise à jour des modèles d'objets en section 2.4, l'automate fini en section 2.5 et l'intégration du contexte 3d de la scène en section 2.6. Nous présentons une évaluation qualitative et quantitative de notre algorithme dans la section 3, et finalement concluons sur une synthèse des principaux résultats ainsi que les perspectives d'évolution dans la partie 4.

2. Méthode

2.1. Aperçu de l'algorithme

Avant de présenter l'algorithme, nous souhaitons préciser la définition des termes suivants :

Un objet : est une modélisation d'une entité réelle dont on souhaite réaliser le suivi ; par exemple, un piéton ou une voiture. A chaque objet nous associons, un identifiant unique (ID), un modèle (voir section 2.4) ainsi qu'un état (voir table 1 et figure 5) qui permet de synthétiser de manière compacte et intelligible l'évolution de l'objet.

Un blob : est une composante connexe issue du masque de mouvement.

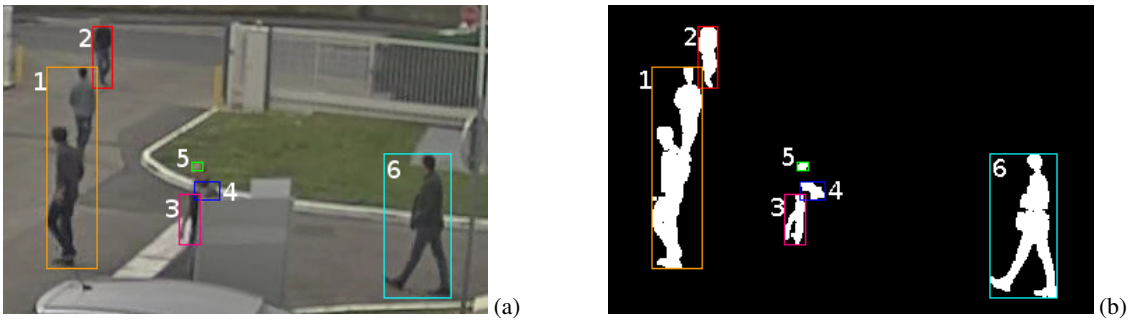
Un groupe : est une modélisation d'un ensemble d'entités réelles qui ne sont plus suivies individuellement, mais collectivement. Comme pour les objets, le groupe dispose d'un ID, d'un modèle et d'un état, mais aussi des identifiants des objets qu'il contient. Ceci permet de réassigner correctement, les identités des objets lorsque le groupe se sépare.

L'algorithme de suivi (figure 2) comporte quatre étapes : La première consiste à extraire les blobs du masque de mouvement. Pour chaque composante connexe extraite du masque de mouvement, nous calculons également un ensemble de caractéristiques de position (centroïde, position réelle au sol), de forme (boite englobante, aire, taille réelle) et d'apparence (histogramme de couleur RGB), qui serviront à évaluer la similarité entre un modèle d'objet et un blob.

La deuxième permet d'associer les objets précédemment suivis et les blobs détectés à l'image courante. Le processus d'association est réalisé en trois passes consécutives (détaillées en section 2.3), et s'appuie sur une mesure de similarité explicité en section 2.2. À l'issue de ces trois passes, les objets non associés sont considérés "perdus" et chaque blob non associé donne naissance à un nouvel objet. Cette étape gère, en outre, les cas d'occultation inter-objets, ainsi que les formations et séparations de groupes.

La troisième étape réalise la mise à jour des modèles des objets. Cette mise à jour dépend des associations réalisées à l'étape précédente et de l'état courant de l'objet. La logique de mise à jour est détaillée en section 2.4.

Enfin, la dernière étape met à jour l'état des objets en réalisant les transitions éventuelles conformément à l'automate fini présenté en section 2.5.



(a) (b)

Figure 1: *Exemples de problèmes liés à l'emploi d'un masque de mouvement : Le blob 1 correspond à un groupe de 2 objets, l'un occultant l'autre. Les blobs (3,4,5) correspondent à des morceaux d'un même objet. Les blobs 2 et 6 correspondent bien à des piétons correctement segmentés.*

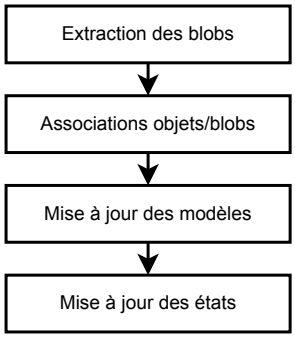


Figure 2: *Grandes étapes de l'algorithme de suivi*

2.2. Calcul des similarités objets/blobs

Pour évaluer la similarité entre l'objet i et le blob j, nous utilisons quatre critères de similarité, chacun donnant une valeur comprise entre 0 (pas similaires) et 1 (identiques). À noter également, que nous compensons le retard temporel des objets sur les blobs (les objets suivis relèvent de l'image à $t-1$, alors que les blobs sont des détections à l'instant t), en prédisant les modèles des objets à l'instant t à partir du modèle à l'instant $t-1$ grâce à un filtre de Kalman (plus de détail en section 2.4).

Le premier critère porte sur la distance entre centroïdes de l'objet i et du blob j :

$$s_{ij}^p = \begin{cases} 1 - d_{ij}/d_{\max} & \text{si } d_{ij} \leq d_{\max} \\ 0 & \text{sinon.} \end{cases} \quad (1)$$

où $d_{\max}$ est la distance maximale admissible entre objets et blobs, et d_{ij} est la distance entre le centroïde prédit de l'objet i et celui du blob j. Ainsi si l'objet i et le blob j sont proches, s_{ij}^p tendra vers 1, et si ils sont lointains, s_{ij}^p tendra vers 0.

Le deuxième critère porte sur la forme (aire et hauteur réelle) de l'objet i et du blob j :

$$s_{ij}^s = 1 - \sqrt{\frac{(\Delta A_{ij})^2 + (\Delta h_{ij})^2}{2}} \quad (2)$$

où ΔA_{ij} et Δh_{ij} représentent la différence relative entre l'aire (resp. la hauteur réelle) de l'objet i et du blob j. Ainsi si les aires et hauteurs réelles de l'objet i et du blob j sont proches, leur différence relative sera nulle et donc s_{ij}^s sera proche de 1.

Le troisième critère porte sur l'apparence. Plus précisément, il mesure le recouvrement entre les histogrammes de couleur de l'objet i et du blob j de la manière suivante :

$$s_{ij}^a = \mathrm{BC}_{i,j} = \frac{1}{B}\sum_{b=1}^{B}\sqrt{h_i(b)h_j(b)} \quad (3)$$

Ici, B est le nombre de classes d'un histogramme de couleur, et $\mathrm{BC}_{i,j}$ est le coefficient de Bhattacharyya entre l'histogramme de couleur h_i du modèle d'objet i, et l'histogramme de couleur h_j du blob j. Pour des histogrammes présentant un fort recouvrement, s_{ij}^a tendra vers 1. En revanche, pour des histogrammes n'ayant pas un bon recouvrement s_{ij}^a tendra vers 0.

Le dernier critère porte sur le recouvrement de la silhouette prédite du cuboïde représentant l'objet i, et le blob j :

$$s_{ij}^o = \begin{cases} 1 & \text{si la silhouette prédite du cuboïde de l'objet } i \\ & \text{intersecte le blob } j \\ 0 & \text{sinon.} \end{cases} \quad (4)$$

A l'inverse des autres critères, qui varient continument entre 0 et 1, le critère de recouvrement est binaire, et ne prend donc que les valeurs 0 ou 1. Il permet de favoriser la détection de la création et de la destruction des groupes.

Finalement, nous définissons la similarité entre un objet i et un blob j par

$$s_{ij} = \sqrt{\frac{\alpha_p(s_{ij}^p)^2 + \alpha_s(s_{ij}^s)^2 + \alpha_a(s_i^a j)^2 + \alpha_o(s_{ij}^o)^2}{\alpha_p + \alpha_s + \alpha_a + \alpha_o}} \quad (5)$$

Les coefficients α définissent les poids de chacun des critères de similarité. Nous détaillerons les valeurs choisies pour ces poids dans la section 2.3. Par construction, s_{ij} est compris entre 0 et 1 et mesure la similarité entre l'objet i et le blob j.

2.3. Association objets/blobs

L'algorithme d'association prend en compte l'état des objets à associer afin de spécialiser les traitements. Aussi nous

invitons le lecteur à se familiariser avec les différents états, en lisant la description de ceux-ci en section 2.5, avant de poursuivre la présentation de l'algorithme d'association.

Par ailleurs, l'algorithme d'association objets/blobs utilise la notion de matrice de similarité S, dont les éléments s_{ij} sont les indices de similarité entre l'objet o_i et le blob b_j, tels que définis dans la section 2.2. Nous illustrons différents cas d'une telle matrice à la figure 3.

L'algorithme d'association se déroule en trois passes successives, chacune tentant d'associer les objets et blobs n'ayant pas encore été associés. Un aperçu de ces trois passes est donné à la figure 4, et sont décrites dans ce qui suit.

La première passe (notée ϕ_1), consiste à associer les objets et blobs ayant une forte similarité. Durant cette phase, tous les objets hormis ceux dans l'état EN GROUPE ou SUPPRIMÉ sont considérés. Les poids des différents critères de similarité sont choisis ainsi : $(\alpha_p = 1, \alpha_s = 0, \alpha_a = 0, \alpha_o = 0)$ pour les objets NOUVEAU, car pour ceux-ci seul la position est jugée fiable ; $(\alpha_p = 0, \alpha_s = 1, \alpha_a = 1, \alpha_o = 0)$ pour les objets PERDU, car pour eux ni la position ni le recouvrement ne peuvent être considérés fiables ; $(\alpha_p = 1, \alpha_s = 1, \alpha_a = 1, \alpha_o = 0)$ pour les objets SUIVI, STABLE et SORTANT. La stratégie d'association est de type glouton (voir algorithme (b) figure 4), et le seuil minimal d'association T_1 est choisi élevé (ici 0.9).

La deuxième passe (notée ϕ_2), tente d'associer les blobs et objets SUIVI, STABLE ou SORTANT restants en prenant en compte les possibilités d'associations multiples (voir algorithme (c) figure 4). Les poids des différents critères sont les suivants $(\alpha_p = 1, \alpha_s = 0, \alpha_a = 1, \alpha_o = 1)$. Le seuil minimal d'association T_2 est fixé à 0.66. Le raisonnement permettant de prendre en compte les associations multiples est le suivant : une fois la meilleure association sélectionnée (o_i, b_j), on recherche dans la ligne i, d'autres blobs notés $\mathcal{B}$, distinct de b_j dont la similarité avec l'objet o_i est supérieure au seuil T_2. On procède de même pour la colonne j, où l'on recherche d'autres objets $\mathcal{O}$, distinct de o_i dont la similarité avec le blob b_j est supérieure au seuil T_2. On a alors plusieurs cas : Si $\mathcal{B} = \mathcal{O} = \emptyset$, on est dans le cas d'association simple 1 objet, 1 blob : l'association (o_i, b_j) est validée. Si $\mathcal{B} \neq \emptyset$ et $\mathcal{O} = \emptyset$, on est dans le cas où l'objet o_i peut être associé à plusieurs blobs. Ce cas peut correspondre soit à un défaut de segmentation (*ie* où l'objet o_i est morcelé en plusieurs blobs) et dans ce cas il convient de fusionner les blobs correspondants en un seul blob et associer ce dernier avec l'objet o_i ; soit à la fin d'une occultation (*ie* si o_i est un groupe), dans ce cas, le groupe o_i est détruit, et on procède à une association gloutonne simple (algorithme (b) figure 4, avec un seuil $T_g = 0.66$ et des poids $(\alpha_p = 0, \alpha_s = 1, \alpha_a = 1, \alpha_o = 0)$) entre les membres du groupe et les blobs de $\mathcal{B} \cup b_j$. Si $\mathcal{O} \neq \emptyset$ et $\mathcal{B} = \emptyset$, on est dans le cas où le blob b_j peut être associé à plusieurs objets. Nous considérons que ceci marque le début d'une occultation, et dans ce cas il convient de créer un groupe comportant les objets $\mathcal{O} \cup o_i$, et d'associer celui-ci au blob b_j. Si $\mathcal{O} \neq \emptyset$ et $\mathcal{B} \neq \emptyset$, par souci de simplicité, nous négligeons le fait qu'il y ait d'autres blobs que b_j, et nous nous ramenons au cas de création d'un groupe.

Pour la troisième et dernière passe (notée ϕ_3), tous les objets non encore associés, hormis ceux dans l'état EN GROUPE ou SUPPRIMÉ, sont considérés afin d'être associés aux blobs restants. Les poids des différents critères de similarité et la stratégie d'association (algorithme (b) figure 4) sont les mêmes que pour la première passe (ϕ_1), seul le seuil minimal d'association change : il est choisi plus bas ($T_3 = 0.75$) afin de permettre d'effectuer les associations qui n'auraient pas été faites durant les deux passes précédentes.

Enfin, à l'issue de ces trois phases d'associations, pour chaque blob restant (non associé), un nouvel objet est créé. Les objets restants sont marqués comme non associés.

L'intérêt de ce déroulement en trois passes est le suivant : la première passe (ϕ_1) permet de propager les associations relativement sûres, et permet de diminuer le nombre d'objets et blobs en entrée de la seconde passe (ϕ_2), plus complexe et donc plus enclin à faire de mauvaises associations. La deuxième passe (ϕ_2) gère les association multiples, et utilise le critère de recouvrement au lieu du critère de forme afin de mieux détecter les formations et destruction de groupes. La troisième passe (ϕ_3) tente d'associer les objets restants, quitte à utiliser un seuil d'association moindre, afin d'éviter la création erronée de nouveaux objets ou le marquage erroné d'objets comme non associé.

2.4. Modèle d'objet : initialisation et mise à jour

Un modèle d'objet comporte les caractéristiques suivantes :

- **position et taille 2d :** un centroïde (x, y), des dimensions de boite englobante 2d (w, h), et une aire A.
- **position, orientation et taille 3d :** un cuboïde 3d posé sur le sol à la position $(x_c, y_c, 0)$, de dimensions (w_c, d_c, h_c) dirigé dans la direction (θ_z).
- **apparence :** histogramme de couleur RGB quantifié. Nous utilisons un histogramme 3d, et quantifions les composantes R, G et B sur 3, 3 et 2 bits respectivement.

L'initialisation d'un modèle d'objet est faite à partir des caractéristiques du blob ayant conduit à la création de l'objet. Le cas délicat d'initialisation du cuboïde 3d est non trivial et mérite quelques explications.

Le but est d'estimer les paramètres du cuboïde conduisant à un masque de détection donné. Nous formulons ce problème d'estimation des paramètres du cuboïde, comme un problème d'optimisation d'énergie. Nous appelons $r(x_c, y_c, w_c, d_c, h_c, \theta_z)$ la fonction qui dessine la silhouette du cuboïde défini par les paramètres donnés, telle qu'elle serait vue par la caméra ; m est le masque de mouvement courant ; $R1$ la boite englobante de la silhouette, $R2$ la boite englobante du blob, et $I_{R1 \cup R2}$ la fonction indicatrice de la région $R1 \cup R2$. Nous définissons alors l'énergie suivante :

$$E(x_c, y_c, w_c, d_c, h_c, m) = \\ ||(r(x_c, y_c, w_c, d_c, h_c, \theta_z) - m).I_{R1 \cup R2}||_{L2} \quad (6)$$

Cette énergie correspond à la distance L2 entre la silhouette prédite du cuboïde et le masque de détection sur la région $R1 \cup R2$.

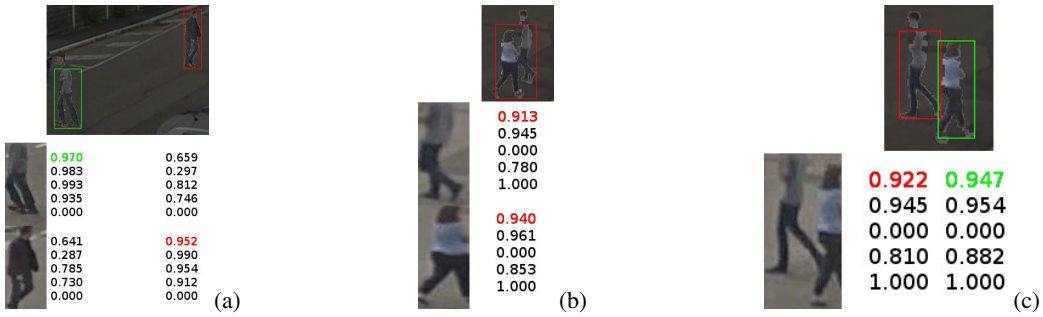


Figure 3: *Illustration des matrices de similarité et de différents cas d'association. Sur les matrices, les modèles d'objets sont sur la première colonne, les blobs détectés sur la première ligne. Un quintuplet de valeurs (s_{ij}, s^p_{ij}, s^s_{ij}, s^a_{ij}, s^o_{ij}) est donné dans cet ordre pour chaque couple (objet i, blob j). Les associations réalisées sont colorées. (a) Cas d'associations simples, chaque objet est associé au blob lui correspondant le mieux. (b) Cas de début d'occultation détecté (ie deux objets correspondent au même blob), un groupe sera donc créé à partir de ces deux objets. (c) Cas de fin d'occultation détecté, (ie un groupe correspond à deux blobs), le groupe sera donc détruit et les objets composant le groupe seront réaffectés aux blobs.*

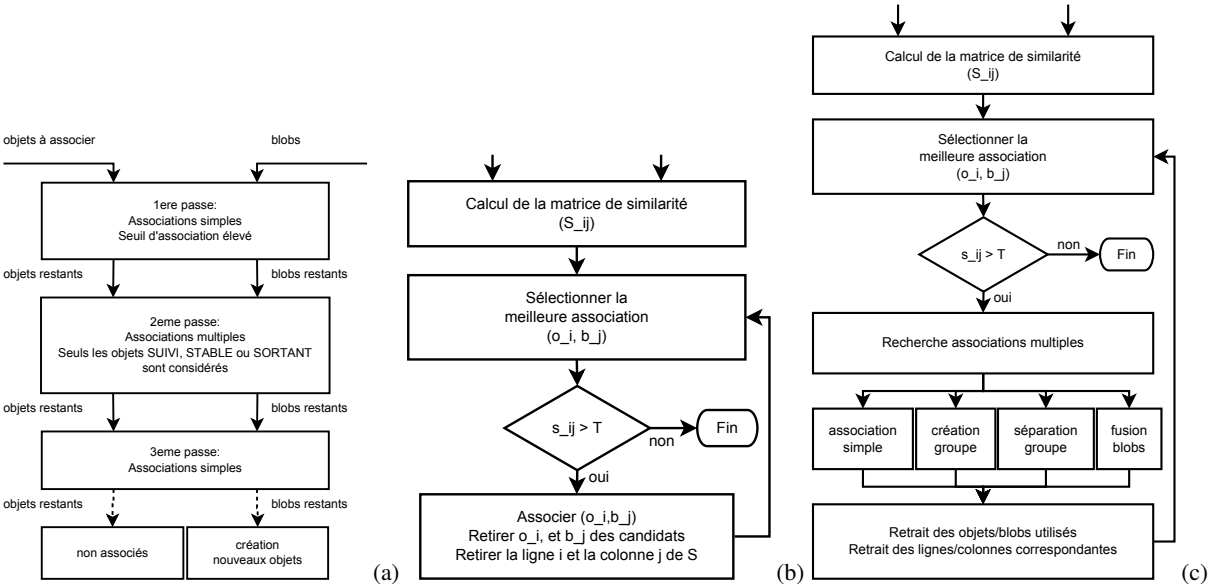


Figure 4: *Algorithmes d'association : (a) : aperçu ; (b) : algorithme glouton pour les associations simples ; (c) algorithme glouton pour les associations multiples*

L'estimation des paramètres du cuboïde est alors réalisée via la minimisation suivante :

$$[x_c, y_c, w_c, d_c, h_c, \theta_z]_0 = \text{argmin}(E(x_c, y_c, w_c, d_c, h_c, \theta_z)) \quad (7)$$

Ne disposant pas des informations de gradient d'énergie pour guider l'optimisation, nous utilisons la méthode de Nelder-Mead [NM65], qui ne requière que les valeurs de la fonction (pas ses dérivées). Cette méthode d'optimisation itérative, utilise la notion de simplexe, que l'on déforme, déplace et réduit progressivement afin que ses sommets se rapprochent d'un point où la fonction est minimale. Cette approche est heuristique, et ne garantie pas la convergence vers un minimum global, c'est pourquoi, nous adoptons la stratégie donnée par l'algorithme 1 pour pallier ces inconvénients et garantir un temps d'exécution borné ($N \times M$ itérations).

Pour faciliter l'estimation initiale, nous réduisons le nombre de paramètres à optimiser en forçant une orientation $\theta_z = 0$, et en imposant une base carrée pour le cuboïde ($w_c = d_c$). Ces contraintes sont relâchées pour les estimations ultérieures afin de permettre au système de converger vers une solution plus proche de la réalité.

Comme nous l'avons déjà évoqué dans la partie 2.2, nous utilisons un filtre de Kalman pour lisser les tailles et positions de la boite englobante 2d et 3d des objets suivis. Plus précisément, nous faisons l'hypothèse de déplacement à vitesse constante, ainsi le vecteur d'état du filtre de Kalman est le suivant :

$$[x, y, \dot{x}, \dot{y}, w, h, \dot{w}, \dot{h}, x_c, y_c, \dot{x_c}, \dot{y_c}, w_c, d_c, h_c, \dot{w_c}, \dot{d_c}, \dot{h_c}, \theta_z] \quad (8)$$

où les quantités marquées d'un point représentent les dérivées instantanées des quantités sans point correspondantes.

Listing 1: *Algorithme d'optimisation permettant d'améliorer la convergence vers un minimum global. Partant d'une estimation initiale x_0 l'algorithme réalise N essais de M itérations de l'algorithme de Nelder-Mead et renvoie la meilleure estimation x_{best}. Chacun de ces essais part d'une position aléatoire autour de la dernière meilleure estimation x_{best}.*

```
x_best = run_optimisation(x_0) {
  x_best = x_0;
  for(i = 1; i < N; ++i) {
    // initialise un simplexe X
    // aléatoirement autour
    // du point x_best
    X = init_simplex_around(x_best);

    // exécute M itérations
    // de Nelder-Mead
    for(j = 1; j < M; ++j)
      X = nelder_mead(X);

    // stocke le centre x du simplexe
    x = barycentre(X);

    // met à jour le meilleur état
    if (cost(x) < cost(x_best))
      x_best = x;
  }
  return x_best;
}
```

Id	Description condition de transition
c_1	l'objet est complètement dans la scène
c_2	l'objet n'est pas complètement dans la scène
c_3	l'objet n'a pas été associé à un blob
c_4	l'objet a disparu depuis trop longtemps ($> T_f$)
c_5	l'objet a été associé
c_6	l'objet entre dans un groupe
c_7	l'objet sort d'un groupe
c_8	l'objet a été associé durant la 1repasse
c_9	l'objet n'a pas été associé durant la 1repasse

Table 1: *Description des conditions de transition.*

2.5. Évolution d'un objet et Automate Fini

L'évolution d'un objet est modélisée à l'aide de l'automate fini représenté figure 5. La table 1 présente les significations de chacun des états et des conditions de transition.

Nous détaillons, à présent, les états considérés ainsi que les transitions possibles entre ces états :

NOUVEAU (s_0) : c'est l'état initial de tout objet d'intérêt. Cet état correspond au cas où l'objet vient d'être créé, mais n'est pas encore entièrement rentré dans la scène (par exemple sur un bord de la scène). Si l'objet entre entièrement dans la scène (c_1), il passe alors dans l'état SUIVI (s_1). En revanche, si l'objet disparait de la scène

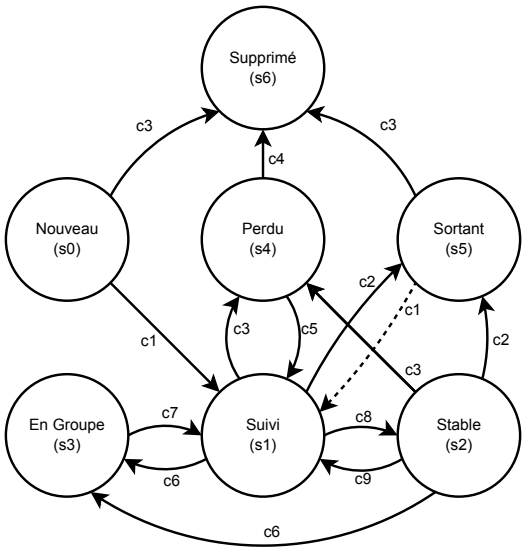


Figure 5: *Schématisation de l'automate fini utilisé. L'état initial est* NOUVEAU *(s_0), l'état final est* SUPPRIMÉ *(s_6). Les conditions de transition notées de c_1 à c_9 sont décrites dans le tableau 1. La transition marquée en pointillés a été ajoutée par rapport à l'article original.*

(c_3), il passe dans l'état SUPPRIMÉ (s_6). L'intérêt de l'état NOUVEAU (s_0) est d'une part de retenir provisoirement les objets qui sont en train de rentrer dans la scène, sans pour autant être totalement visible, et d'autre part, de permettre de filtrer les détections ponctuelles insignifiantes qui ne passeront jamais à l'état SUIVI (s_1).

SUIVI (s_1) : l'objet est entièrement dans la scène, mais sa stabilité n'est pas encore établie. Si l'objet est entré dans un groupe (c_6), celui-ci passe à l'état EN GROUPE (s_3). Si l'objet n'a pas été associé à un blob (c_3), celui-ci passe à l'état PERDU (s_4). Si l'objet n'est plus entièrement dans la scène (c_2), il passe à l'état SORTANT (s_5). Si l'objet est associé durant la première phase d'association (c_8), il passe à l'état STABLE (s_2).

STABLE (s_2) : l'objet est entièrement dans la scène et est considéré stable. Si l'objet est entré dans un groupe (c_6), celui-ci passe à l'état EN GROUPE (s_3). Si l'objet n'a pas été associé à un blob (c_3), celui-ci passe à l'état PERDU (s_4). Si l'objet n'est plus entièrement dans la scène (c_2), il passe à l'état SORTANT (s_5). Si l'objet n'est pas associé durant la première phase d'association (c_9), il passe à l'état SUIVI (s_1). L'utilité de l'état STABLE (s_2) est de traduire, pour les éventuels traitements de plus haut niveau que le suivi, un degré de confiance accru par rapport aux objets SUIVI (s_1).

EN GROUPE (s_3) : l'objet n'est plus suivi individuellement, seul le groupe auquel il appartient est suivi. Cet état permet de conserver les objets qui ne peuvent plus être suivi individuellement afin de permettre de réaffecter correctement ceux-ci lorsque le groupe se scindera (c_7). Dans ce cas, l'objet retourne à l'état SUIVI (s_1).

PERDU (s_4) : l'objet n'a pas été associé, par exemple à cause d'une occultation par un obstacle ou une mauvaise

détection. Cependant, son modèle est maintenu en mémoire, afin de permettre de le retrouver s'il venait à réapparaitre. Si l'objet n'est pas réapparu dans la scène depuis trop longtemps (c_4), c'est à dire, non associé consécutivement plus de T_f fois, il passe alors à SUPPRIMÉ (s_6). Si l'objet est retrouvé (c_5), il passe à l'état SUIVI (s_1).

SORTANT (s_5) : l'objet est en train de sortir de la scène (par exemple il touche le bord). Si l'objet n'est plus détecté (c_3), il passe à l'état SUPPRIMÉ (s_6). Si l'objet re-rentre dans la scène (c_1), il passe alors dans l'état SUIVI (s_1). À la différence de l'état PERDU (s_4), qui traduit la perte d'un objet généralement due à un problème de détection, l'état SORTANT (s_5) traduit l'intention d'un objet suivi de sortir de la scène. Il faut également le distinguer de l'état NOUVEAU (s_0), car bien qu'à la bordure de la scène considérée, l'objet SORTANT (s_5) est entré dans la scène et manifeste la volonté d'en sortir.

SUPPRIMÉ (s_6) : l'objet n'est plus suivi et peut être oublié. C'est l'état final de tout objet d'intérêt.

Si pour un état donné, aucune des conditions de transition ne s'applique, l'objet garde son état.

2.6. Intégration du contexte 3d

L'utilisation d'une caméra calibrée rend possible un ensemble de raisonnements tridimensionnels que nous allons détailler. À noter cependant, que comme seule une seule caméra est utilisée, il n'est pas possible de résoudre avec certitude l'ambiguïté de profondeur sans apport d'informations supplémentaires.

Les critères de transition c_1 et c_2 nécessitent de définir la zone correspondant à la scène surveillée, ainsi que le critère d'appartenance ou non à celle-ci. Étant donné que nous faisons l'hypothèse que les objets se déplacent sur le sol, il est naturel, de considérer qu'un objet est dans la scène quand sa boite englobante est incluse dans l'image, et que le bas de l'objet touche le sol. Cette définition permet d'éviter le suivi des éventuels blobs inintéressants se déplaçant dans le ciel par exemple. Pour générer un masque de sol automatiquement, nous utilisons l'approche proposée par [RTR13] qui permet de faire des rendus synthétiques de vues caméras, prenant en compte les bâtiments environnants grâce à une coordonnée GPS, la base de données OpenStreetMap et la calibration de la caméra. En choisissant les couleurs de rendu de manière appropriée (le sol en blanc, tout le reste en noir), nous obtenons ainsi un moyen simple et automatique d'obtenir un masque de sol prenant en compte les occultations liées aux bâtiments environnants (figure 6).

Par ailleurs, pour calculer les silhouettes des cuboïdes prédits nécessaires pour le calcul du critère de recouvrement s^o_{ij} (équation 4), nous réalisons un rendu synthétique des cuboïdes dans la scène tels qu'ils sont vus par la caméra, chaque cuboïde ayant une couleur unique. Ce rendu tient compte des occultations inter-objets et de celles liées à la scène.

Finalement nous attirons l'attention sur le fait que cette formulation tridimensionnelle a été mise en place dans le but de pouvoir également filtrer les ombres projetées des objets. Cependant, par manque de temps, ce filtrage n'a pas été mis en place. L'idée est d'intégrer le modèle de prédiction d'ombres projetées développé par [RTR13] dans la phase d'estimation du cuboïde via l'ajout d'une variable booléenne présence/absence de l'ombre dans la fonction de rendu. L'optimisation de l'énergie sera alors capable de déduire si la présence de l'ombre est compatible avec le masque de mouvement, et dans ce cas la filtrer.

3. Évaluation

3.1. Éléments d'implémentation

Nous avons implémenté en C++ l'algorithme décrit dans [LFP*13]. Cette implémentation, que nous espérons fidèle à la méthode originale, constitue notre référence, et nous permet d'évaluer l'apport de nos modifications pour le rendre pleinement utilisable dans le contexte de la vidéosurveillance.

Les masques de mouvement sont obtenus à l'aide du modèle de fond ViBe [BVD11], suivi d'opérations morphologiques filtrant les pixels isolés.

Le seuil d'association est fixé à 0.66 pour la méthode originale, et pour notre méthode, nous utilisons les valeurs fixes suivantes : $T_1 = 0.9, T_2 = T_g = 0.66, T_3 = 0.75$. De la même manière que pour la méthode originale, nous fixons T_f de manière à oublier les objets perdus depuis plus de 1 seconde.

3.2. Validation qualitative

Afin de valider l'approche proposée, nous avons réalisé un ensemble de vidéos synthétiques courtes à l'aide du logiciel de création et d'animation 3d Muvizu [†]. Chaque séquence synthétique a été réalisée afin de tester un aspect précis d'un algorithme de suivi. La première vidéo montre deux piétons marchant vers la caméra ; à mi-chemin l'un passe devant l'autre, puis avant d'arriver au pied de la caméra, ils repartent chacun de leur côté. La deuxième vidéo montre le croisement d'un piéton et d'une voiture ; le piéton vient de la gauche, la voiture de la droite. Nous avons également utilisé la vidéo de test de la séquence 1 du challenge PETS 2001 [‡], qui met en scène des piétons et des voitures. Nous présentons dans ce qui suit quelques exemples de résultats montrant les apports de notre méthode. Une validation qualitative sur un autre jeu de données (pour lesquelles nous disposons de la vérité terrain) est présenté dans la section 3.3.

La première séquence (lignes 1 et 2 de la figure 7) montre la difficulté qu'a l'algorithme original (ligne 2) à détecter les débuts d'occultation. On note cependant que grâce à l'état PERDU, l'algorithme retrouve et réassocie correctement le piéton qui a été masqué totalement. En revanche, pour notre version (ligne 1), la détection du début d'occulation, la création du groupe associé et la fin d'occultation ont été correctement réalisées.

La deuxième séquence (lignes 3 et 4 de la figure 7) montre un cas où l'algorithme original (ligne 4) n'arrive pas à détecter le début d'occultation (à cause de la différence de taille

†. http ://muvizu.com
‡. http ://www.cvg.reading.ac.uk/PETS2001

Figure 6: *Exemple de génération du masque de sol utilisant le framework de rendu de [RTR13] : (a) scène originale, (b) carte sémantique synthétique (sol en rouge, ciel en noir, bâtiments en vert), (c) masque de sol*

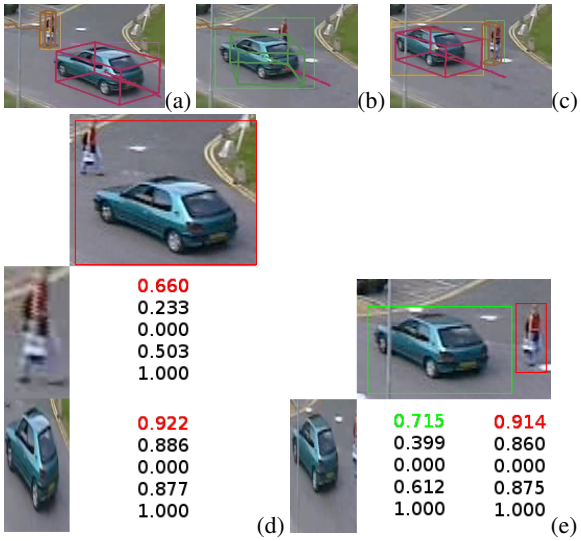


Figure 8: *Cas d'occultation d'un piéton par une voiture (a,b,c) et matrices d'associations ayant permis la détection du début (d) et de la fin (e) d'occultation. La fusion du blob du piéton et de la voiture est détectée (d), un groupe est alors créé et associé au blob regroupant le piéton et la voiture (b). À la scission de ce blob (e), les modèles individuels du piéton de la voiture sont correctement réassociés (c).*

des objets) et cause donc une perte, puis la création d'un nouvel objet. Notre algorithme (ligne 3), quant à lui, détecte correctement le début et la fin de l'occultation, permettant ainsi de suivre correctement la voiture et le piéton.

La séquence de test 1 du challenge PETS 2001 confirme les observations faites sur les séquences synthétiques : notre algorithme améliore la détection de début de fin d'occultation. Une illustration plus précise d'occultation entre un piéton et une voiture est donné à la figure 8.

3.3. Validation quantitative

Pour l'évaluation quantitative, nous avons utilisé deux vidéos réelles : PETS09.S2L1, qui fait partie du challenge PETS [§] ; et une vidéo (parking1) capturée sur un parking proche de nos locaux. Les caractéristiques principales des

Vidéo	Résolution	# Img	Img/s	# objets
PETS.S2L1	768x576	795	7	19
parking1	1280x960	3600	25	27

Table 2: *Caractéristiques des séquences vidéos de test*

vidéos (résolution, nombre d'images, nombre d'images par seconde et nombre d'objets) utilisées pour la validation sont répertoriées dans le tableau 2. Pour la vidéo PETS09.S2L1, les données de calibrations sont fournies et ont donc été employées. Les coordonnées GPS permettant de retrouver les bâtiments environnants sont également utilisées. Pour la vidéo (parking1), nous avons réalisé une calibration sommaire à l'aide d'un logiciel interne de simulation et de placement de caméras.

Pour la vidéo PETS.S2L1, nous utilisons la vérité terrain mise à disposition par A. Milan [¶]. Pour la vidéo parking1, nous avons étiqueté manuellement la séquence.

Nous utilisons les métriques d'évaluation CLEARMOT décritent dans [BS08], en particulier les mesures : *Multiple Object Tracking Accuracy* (MOTA), *Multiple Object Tracking Precision* (MOTP), *False Positive* (FP), *False Negative* (FN), et *ID Switch* (IDs). Nous classons également les trajectoires en terme de *Mostly Tracked* (MT), *Partially Tracked* (PT) et *Mostly Lost* (ML) selon le pourcentage de la trajectoire effectivement suivi [LHN09] : si 80% ou plus de la trajectoire d'un objet est correctement suivie, la trajectoire est considérée MT, si moins de 20% de la trajectoire d'un objet est correctement suivie, la trajectoire est considérée ML, sinon elle est considérée PT.

Avant toute remarque, nous souhaitons attirer l'attention sur la difficulté d'évaluer objectivement un algorithme de suivi étant donné la dépendance à la qualité des détections, la qualité de la vérité terrain, ou l'implémentation de l'algorithme d'évaluation [MSR13]. Nous précisons également que nous exportons les groupes et non les objets le composant en cas d'occultations ; ainsi une certaine baisse de performance est observée car la vérité terrain considère les objets séparément et non le groupe.

L'algorithme proposé améliore sensiblement l'algorithme initial (table 3), et semble mieux gérer les cas de début

§. http ://www.cvg.rdg.ac.uk/PETS2009/

¶. `http://www.milanton.de/data.html`

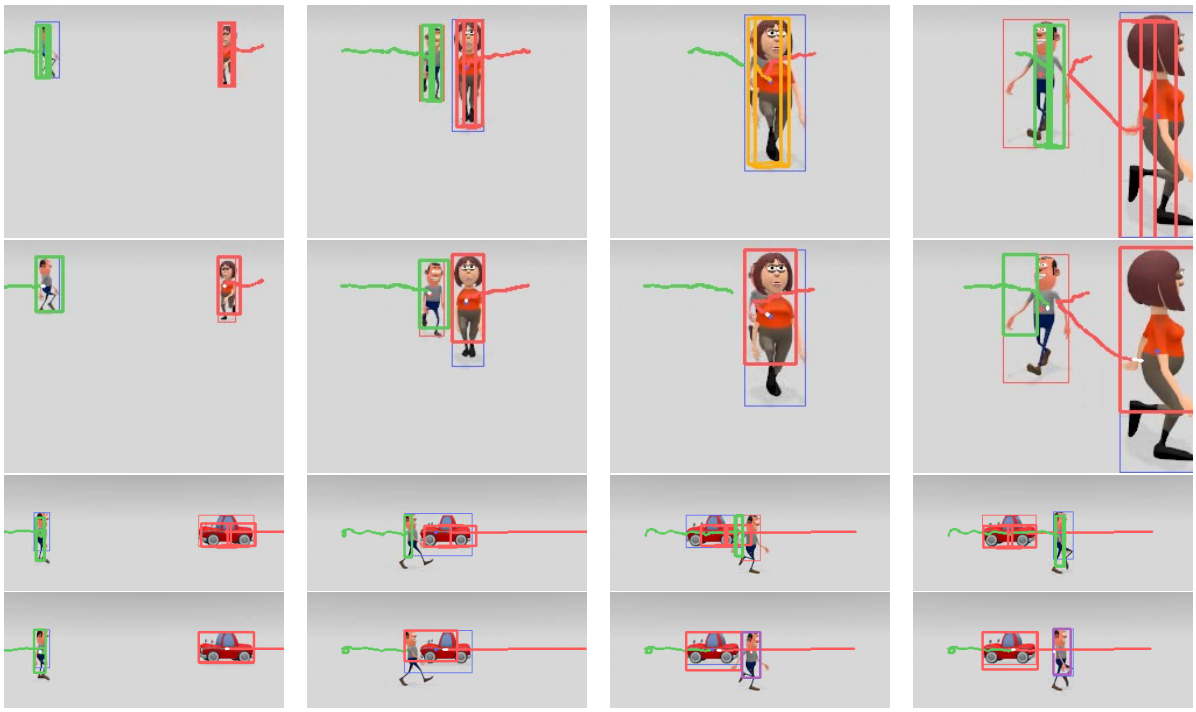


Figure 7: *Illustration sur séquences synthétiques, et comportement des algorithmes sur deux cas d'occultations (piéton/piéton et piéton/voiture). La première (resp. deuxième) ligne montre le comportement de notre algorithme (resp. l'algorithme original) sur la première séquence synthétique (cas piéton/piéton). La troisième (resp. quatrième) ligne montre le comportement de notre algorithme (resp. l'algorithme original) sur la deuxième séquence synthétique (cas piéton/voiture).*

Video	Methode	MOTA	MOTP	GT	FP	FN	IDs	Précision	Recall	MT	PT	ML
PETS.S2L1	[LFP*13]	0.759	0.594	4644	176	762	179	0.957	0.836	13	6	0
	Notre Méthode	0.774	0.617	4644	114	737	199	0.972	0.841	14	5	0
parking1	[LFP*13]	0.705	0.510	15439	226	4196	135	0.980	0.728	12	15	0
	Notre Méthode	0.729	0.516	15439	200	3754	228	0.983	0.757	12	15	0

Table 3: *Performance des algorithmes de suivi*

et fin d'occultation (figure 9). Quasiment tous les indicateurs montrent une amélioration, à l'exception du nombre d'échanges d'identité (IDs). Une des causes expliquant le nombre plus important d'échanges d'identité concerne la destruction des groupes : dans les cas où notre algorithme détecte la séparation d'un groupe alors que l'algorithme original ne le détecte pas (par exemple à la figure 9), la réaffectation d'identité à la destruction du groupe causera potentiellement un échange d'identité supplémentaire (ID groupe vers ID objet), que l'algorithme original ne fera pas. Par ailleurs, on notera la baisse du nombre de faux positifs (FP) et du nombre de faux négatifs (FN) qui sont appréciables dans un contexte de vidéosurveillance : les omissions (faux négatifs) ne sont pas tolérables car elles sont contraires aux objectifs d'un système de détection d'intrusion ; alors que les fausses alarmes (faux positifs) tendent à réduire la confiance des opérateurs en leur système de surveillance.

Il est à noter que notre méthode n'est pas encore capable d'améliorer les cas où les objets allant former un groupe sont de taille très inégale (figure 10) : dans ce cas, l'objet de dimension prépondérante risque d'être associé dès la première

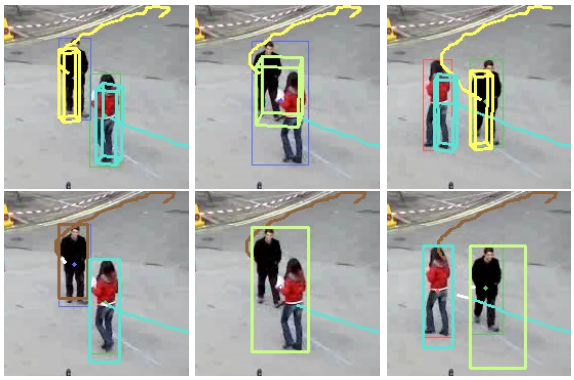


Figure 9: *Exemple de création et destruction de groupe mieux géré par notre approche : dans notre cas, le groupe est créé puis détruit correctement. Alors que dans l'algorithme original, un nouvel objet est créé, et il vole l'identité d'un des piétons.*

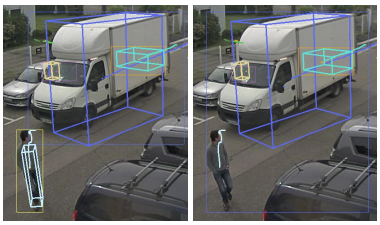


Figure 10: *Cas où les objets sont de dimensions trop inégales, et où la création de groupe n'a pas pu être détectée : Le piéton est "absorbé" par le camion.*

passe (le petit objet est en fait négligeable), ce qui empêche la création du groupe.

Par ailleurs, l'estimation des cuboïdes des objets suivis est délicate et requière un masque de mouvement peu bruité où les pieds/points de contact au sol soient visibles. De plus elle ralentit l'algorithme du fait du nombre d'itérations et de la formulation de l'énergie ; cependant, cela autorise un raisonnement en 3d que nous souhaitons utiliser afin de détecter et filtrer les ombres présentes sur le masque de mouvement (l'estimation des ombres projetées peut être réalisée comme indiqué dans [RTR13]).

Notons que l'utilisation de la 3d (bâtiments environnants, masque de sol, estimation des cuboïdes) nécessite une caméra calibrée, cependant, pour une application en vidéosurveillance, cette calibration, bien que grossière, peu facilement être obtenue à partir des études de prédéploiement des caméras, ce qui n'est donc pas un frein à son utilisation. Nous l'avons montré sur les vidéos synthétiques et parking1 où une calibration grossière a été effectuée à l'aide du logiciel interne de simulation et de placement de caméras, et non par une procédure de calibration classique.

4. Conclusion

Nous avons proposé un algorithme de suivi multi-objets adapté au contexte de la vidéosurveillance. Celui-ci étend les travaux de [LFP*13] en généralisant le suivi, en améliorant les mécanismes de formation et destruction des groupes, et en intégrant la 3d dans les raisonnements. Nous avons montré l'impact de notre contribution par rapport à l'algorithme initial à la fois sur des exemples synthétiques et sur des exemples réels. En terme de travaux futurs, nous souhaitons étendre le raisonnement 3d afin de prédire et filtrer les éventuelles ombres présentent dans le masque de mouvement.

Références

[BRL*09] BREITENSTEIN M., REICHLIN F., LEIBE B., KOLLER-MEIER E., VAN GOOL L. : Robust tracking-by-detection using a detector confidence particle filter. In *Computer Vision, 2009 IEEE 12th International Conference on* (Sept 2009), pp. 1515–1522.

[BS08] BERNARDIN K., STIEFELHAGEN R. : Evaluating multiple object tracking performance : The clear mot metrics. *J. Image Video Process.. Vol. 2008* (janvier 2008), 1 :1–1 :10.

[BVD11] BARNICH O., VAN DROOGENBROECK M. : Vibe : A universal background subtraction algorithm for video sequences. *Image Processing, IEEE Transactions on. Vol. 20*, Num. 6 (2011), 1709–1724.

[CFPV10] CONTE D., FOGGIA P., PERCANNELLA G., VENTO M. : Performance evaluation of a people tracking system on pets2009 database. In *Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance* (Washington, DC, USA, 2010), AVSS '10, IEEE Computer Society, pp. 119–126.

[FJ14] FÜHR G., JUNG C. R. : Combining patch matching and detection for robust pedestrian tracking in monocular calibrated cameras. *Pattern Recognition Letters. Vol. 39*, Num. 0 (2014), 11 – 20. Advances in Pattern Recognition and Computer Vision.

[GMG12] GODBEHERE A., MATSUKAWA A., GOLDBERG K. : Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In *American Control Conference (ACC), 2012* (June 2012), pp. 4305–4312.

[LFP*13] LASCIO R. D., FOGGIA P., PERCANNELLA G., SAGGESE A., VENTO M. : A real time algorithm for people tracking using contextual reasoning. *Computer Vision and Image Understanding. Vol. 117*, Num. 8 (2013), 892–908.

[LFSV12] LASCIO R. D., FOGGIA P., SAGGESE A., VENTO M. : Tracking interacting objects in complex situations by using contextual reasoning. In *VISAPP (2)* (2012), Csurka G., Braz J., (Eds.), SciTePress, pp. 104–113.

[LHN09] LI Y., HUANG C., NEVATIA R. : Learning to associate : Hybridboosted multi-target tracker for crowded scene. In *CVPR'09* (2009), pp. 2953–2960.

[MSR13] MILAN A., SCHINDLER K., ROTH S. : Challenges of ground truth evaluation of multi-target tracking. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on* (June 2013), pp. 735–742.

[NM65] NELDER J. A., MEAD R. : A simplex method for function minimization. *The Computer Journal. Vol. 7*, Num. 4 (1965), 308–313.

[PA10] PAPADOURAKIS V., ARGYROS A. : Multiple objects tracking in the presence of long-term occlusions. *Comput. Vis. Image Underst.. Vol. 114*, Num. 7 (juillet 2010), 835–846.

[RKDL12] RAHEJA J. L., KALITA S., DUTTA P. J., LOVENDRA S. : A robust real time people tracking and counting incorporating shadow detection and removal. *International Journal of Computer Applications. Vol. 46*, Num. 4 (May 2012), 51–58. Published by Foundation of Computer Science, New York, USA.

[RTR13] ROGEZ M., TOUGNE L., ROBINAULT L. : A Prior-Knowledge Based Casted Shadows Prediction Model Featuring OpenStreetMap Data. In *VISAPP* (février 2013), pp. 602–607.