

Analyse de séquences non calibrées pour la reconstruction 3D de scène

Uncalibrated image sequence analysis for 3D scene reconstruction

Lionel Oisel¹, François Fleuret², Patrick Horain³, Luce Morin¹, Jean-Marc Vezien²
Françoise Prêteux³, André Gagalowicz², Claude Labit¹,
et Pascal Leray⁴

¹ Projet Temis, IRISA/INRIA-Rennes/Université de Rennes 1

² Projet Syntim, INRIA-Rocquencourt

³ Département Sim, Institut National des Télécommunications - Evry

⁴ CNET/DIH/ATI/I3VN

contact : Luce Morin, Irisa, Campus de Beaulieu, 35042 Rennes cedex, luce.morin@irisa.fr

Résumé

Notre objectif est de restituer une représentation 3D synthétisable d'une scène dynamique à partir d'une séquence d'images non calibrées de la scène. Celle-ci comprend quelques objets pour lesquels on dispose d'un modèle générique 3D déformable et qui sont placés dans un environnement quelconque. Les zones fixes non connues peuvent être reconstruites par des techniques projectives. Nous présentons la phase d'appariement dense préalable à la reconstruction. Une méthode de reconnaissance par indexation sur les zones en mouvement permet de détecter et d'identifier les objets connus (ici des visages). Ces objets sont reconstruits en 3D par recalage et déformation d'un modèle 3D générique. Ils permettront de plus la reconstruction dans un repère euclidien du reste de la scène. Les résultats sont présentés en particulier sur des images extraites d'une séquence de bureau avec personnage.

Mots Clefs

séquence vidéo, géométrie projective, régularisation, reconnaissance, indexation, arbre de décision, modèle 3D déformable, recalage, morphologie mathématique

Abstract

This paper aims at restituting a 3D representation of a dynamic scene from an uncalibrated video image sequence. The scene is assumed to be composed of objects for which a generic deformable model is available, surrounded by an unknown environment. The unknown fixed areas can be reconstructed using projective techniques. We present the dense matching step that precedes the reconstruction. The known objects (faces in this example)

are detected and identified by applying an indexation based recognition method on the moving areas. These objects are reconstructed in 3D through global positioning and local deformation. They will also be used to recover an Euclidean representation for the remaining of the scene. Results are shown in particular on images extracted from an uncalibrated sequence featuring a person in an indoor office-like environment.

Keywords

image sequence, projective geometry, regularization, recognition, indexing, decision tree, 3D deformable model, registration, mathematical morphology

1 Introduction

Le développement de nouvelles fonctionnalités de communication et manipulation d'images vidéo [8] (dans le contexte de normalisation MPEG-4 notamment [19], [6]), nécessite l'exploration de nouveaux outils d'analyse et de reconstruction d'objets extraits de scènes visuelles réelles. La notion d'objet ici introduite peut correspondre, selon les applications visées, à divers niveaux d'une représentation hiérarchique de modélisation allant de régions 2D segmentées dans l'image [9] (mosaïque vidéo) à la recherche et l'identification d'objets 3D (rigides ou déformables) en passant par une structuration intermédiaire telle que la décomposition d'une scène en "couches" (layers) où les objets sont ordonnés selon des indices de profondeur qualitatifs. Nous nous plaçons dans le cadre de scènes naturelles quelconques, c'est-à-dire sans *a priori* de modélisation géométrique, cinématique ou photométrique, complexes (par exemple, avec mouvements combinés du capteur et des objets) et pour lesquels les pa-

ramètres de calibration du capteur ne sont pas disponibles. L'identification et la reconstruction totale de la géométrie euclidienne de la scène 3D+t, perçue par un capteur non calibré, reste un objectif inaccessible dans sa généralité. Il convient donc de dissocier les tâches de reconstruction partielle (projective) d'une scène 3D, celles de détection de la présence d'un objet dont un modèle générique est connu, et enfin celles de recalage et déformation d'un modèle générique d'objet 3D reconnu sur l'observation de l'objet réel dans la scène. Ces diverses phases d'études sont conçues et réalisées au sein d'un projet commun à plusieurs partenaires¹.

L'originalité de l'approche proposée réside en :

- la prise en compte d'une base de données d'objets 3D constituant des modèles génériques, base qui peut être évolutive au fur et à mesure de la progression dans l'analyse et selon les différentes formes d'apprentissage en fonction du degré d'interactivité. Nous avons considéré, dans une première étape, l'usage de modèles d'images faciales (visages) pour lesquelles des applications futures font d'ores et déjà l'objet de nombreuses expérimentations [25] : vidéophonie améliorée, téléprésence, téléconférence virtuelle, acteurs synthétiques, ... La démarche méthodologique reste cependant générique et peut se décliner pour d'autres classes d'objets.
- la phase d'analyse guidée par les modèles géométriques introduits conduisant à la création de modèles synthétisables. Cette phase d'analyse est le thème central de l'étude présentée ici ; nous la détaillons ci-après selon ces trois grandes fonctions complémentaires que sont i) la détection/localisation par requête d'indexation de l'objet 3D modélisé, ii) la mise en correspondance, au niveau de cet objet identifié, du modèle et des observations image par modèles de déformation continue et enfin, iii) la reconstruction (projective, dans un premier temps) des autres éléments constitutifs de la scène.
- une étape de reconstruction des images à partir d'algorithmes de synthèse provenant des modèles 3D issus de l'analyse. L'objectif de la recherche réalisée en premier lieu ne vise pas l'optimisation de cette dernière phase qui sera donc réalisée par simple re-projection de la texture photométrique initiale.

Le schéma global (figure 1) de l'algorithme d'analyse actuellement développé s'articule donc autour :

- d'une étape de pré-traitement correspondant à la détection des zones "conformes" au mouvement dominant, c'est-à-dire correspondant aux objets fixes

1. La présente étude s'inscrit dans le cadre de conventions de recherche en cours passées entre le CNET et l'INRIA/Projet Temis, Projet Syntim et le Département Sim de l'INT.

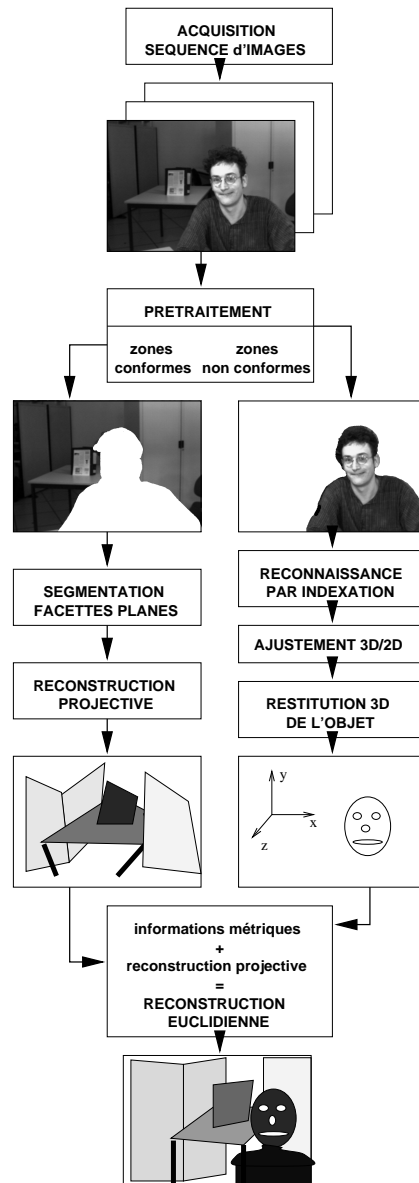


FIG. 1 – Schéma global de l'analyse

dont le mouvement apparent est dû au mouvement de la caméra, et des zones "non conformes" correspondant à des objets mobiles ; cette étape est fondée sur les travaux de J.M. Odobez [18] et n'est pas détaillée ici,

- d'une phase de reconnaissance par indexation,
- d'une méthode d'ajustement d'un modèle 3D sur la séquence avec restitution 3D de l'objet,
- d'une étape d'analyse du mouvement en vue de la reconstruction projective des parties fixes de la scène.

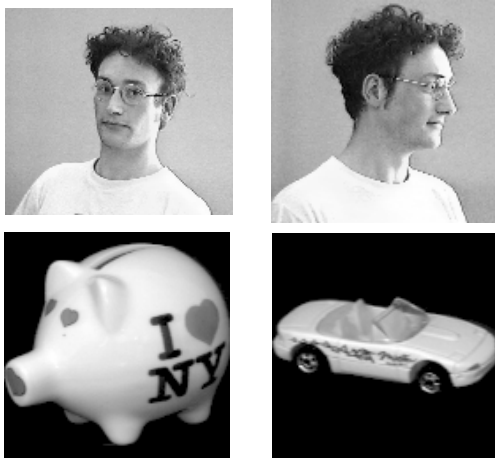


FIG. 2 – Plusieurs images de notre base de données test. En haut, deux images du sujet à identifier, en bas deux autres objets de la base de données Columbia.

2 Détection d'objets connus par indexation

Cette première étape consiste à reconnaître dans la scène d'éventuels objets d'une base de données acquise préalablement par apprentissage. Cela permet d'une part de sélectionner le modèle générique 3D approprié parmi ceux disponibles et d'autre part d'identifier les instants de la séquence pour lesquels il convient de déclencher la procédure d'ajustement 3D et de restitution tridimensionnelle.

La méthode que nous proposons s'inspire d'un algorithme développé par Yali Amit, Donald Geman et Ken Wilder pour la reconnaissance de caractères manuscrits et présenté en détails dans [2]. Elle consiste à construire des arbres de classification qui cherchent, sur l'image traitée, des instances de graphes caractéristiques des objets recherchés. Nous proposons d'utiliser la même technique pour des objets 3D et des images en niveaux de gris.

L'utilisation d'arbres de classification [23] pour des problèmes de reconnaissance de formes n'est pas nouvelle. Cette technique est utilisée pour la classification de formes rigides 2D [2] [26]. Dans [10], les auteurs considèrent des objets 2D déformables et dans [27] des objets 3D rigides.

Dans notre cas, la base de données utilisée pendant l'apprentissage est constituée d'images représentatives de ce que peut contenir la scène, sous forme de vues de chaque objet, isolé, sous de nombreux angles différents (Fig. 2).

2.1 Recodage des images, étiquetage

La première étape de la reconnaissance consiste à recoder l'image afin d'obtenir en chaque point une série de paramètres booléens qui décrivent la topologie locale de l'image. Chacun de ces paramètres décrit une propriété

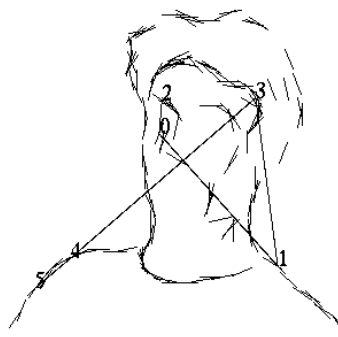


FIG. 3 – Codage utilisant une détection de bords classique, effectué sur la première image de la figure 2. On a superposé les instances d'un graphe simple (profondeur 5) sur cette image.

qui peut être vérifiée ou non. On dira qu'un pixel porte un "tag" de code k si le k -ème paramètre est vrai (un pixel peut donc porter plusieurs tags).

Le codage précédemment utilisé [12] était une adaptation assez directe de ce qui avait été développé pour la reconnaissance de caractères dans [2]. Nous utilisons actuellement un nouveau codage plus proche d'une détection de bords classique (voir Fig. 3), plus adéquat pour rendre la reconnaissance robuste à la présence d'un fond d'image. Chaque pixel porte un seul tag, celui correspondant à la direction du bord présent à cet endroit-là (les directions sont quantifiées en $N = 8$ ou 16 classes, le tag k étant présent si l'orientation du bord est comprise entre $k \cdot \frac{\pi}{N}$ et $(k + 1) \cdot \frac{\pi}{N}$). S'il n'y a pas de bord, le pixel ne porte pas de tag. A ces N tags de bords, nous avons dans certaines expériences rajouté N autres tags qui indiquent la même orientation de manière plus grossière (le tag $N + k$ étant présent si l'orientation du bord est comprise entre $(k - 1) \cdot \frac{\pi}{N}$ et $(k + 2) \cdot \frac{\pi}{N}$). Ces tags plus grossiers ont un pouvoir de séparation plus faible, mais sont plus stables en cas de perturbations de l'image.

2.2 Algorithme de reconnaissance

La reconnaissance proprement dite se fait à l'aide d'un arbre binaire de classification. Chaque nœud de l'arbre porte une "question", c'est-à-dire une configuration géométrique qui peut se trouver ou pas sur une image, configuration qui fait intervenir des propriétés locales et globales. Par exemple, un nœud de l'arbre pourra porter une question du type "a-t-on quelque part dans l'image deux bords verticaux alignés horizontalement?". Dans ce cas, les contraintes locales sont l'existence de bords verticaux (configuration de niveaux de gris ne faisant intervenir qu'une petite zone de l'image), et la contrainte globale est la contrainte d'alignement horizontal. On peut représenter une telle relation (ne faisant intervenir que deux points) avec une arête de graphe: les sommets de cette arête sont les deux points qui vérifient les propriétés locales, et l'arête elle-même représente la contrainte

globale.

Il est à noter que les questions posées sont relatives à des couples de points tels qu'un de ces deux points intervenait déjà dans une question précédente de l'arbre.

La reconnaissance d'une image se fait alors en "posant" successivement les questions portées par les nœuds de l'arbre. En fonction de la réponse donnée à chacune de ces questions, on considère un autre nœud, et on teste une nouvelle question. Un des nœuds fils sera donc associé à une réponse positive, et l'autre à une réponse négative.

L'algorithme se termine quand on arrive à une feuille de l'arbre. Chaque feuille porte des résultats empiriques sur les classes des objets des images d'apprentissage qui ont abouti à cette feuille (par exemple, dans le cas de la base de données Columbia, il y a 20 objets, soit 20 classes, et chaque feuille porte donc une distribution sur $\{0..19\}$). Comme on peut associer à chaque relation une arête de graphe, une série de relations peut être représentée sous la forme d'un graphe. Dans le cas où chaque relation fait intervenir un point, et un seul, d'une autre relation, ce graphe est connexe et sans cycle. Donc, à chaque feuille de l'arbre on peut associer le graphe parcouru par les images qui y aboutissent (Fig. 3).

2.3 Apprentissage

Pour construire les arbres de questions servant à la reconnaissance, on se donne un ensemble d'images étiquetées par leur classe réelle (la classe de l'objet représenté). Nous faisons donc ici un apprentissage supervisé.

La construction des arbres est faite en plaçant à chaque nœud la relation qui permet de réduire le plus l'incertitude sur la classe réelle de l'image que l'on cherche à reconnaître. On évalue empiriquement cette incertitude à l'aide de l'ensemble d'apprentissage.

Pour construire un nœud de l'arbre, on génère aléatoirement un ensemble de plusieurs centaines de relations. Plus cet ensemble est grand, plus le gain d'information pourra être important. Une relation est caractérisée par un triplet (T_1, T_2, α) , et est associée à la question "existe-t-il quelque part sur l'image un point x portant le tag T_1 et un point y portant le tag T_2 , tel que la droite (x, y) fasse un angle α avec l'horizontale?". Pour chacune de ces relations, on évalue le gain d'information obtenu à l'aide de l'entropie de Shannon (que l'on minimise) [1]. La procédure est itérée au fur et à mesure de la construction de l'arbre, et se termine quand toutes les images qui aboutissent à un nœud sont d'une classe unique.

Un point crucial de cet algorithme réside dans le fait que la construction d'un arbre dépend du choix des ensembles de questions aléatoires utilisés lors de la construction de chaque nœud. En utilisant des ensembles différents, l'algorithme construit des arbres différents. Ainsi, on peut construire plusieurs arbres, et fusionner les différents résultats en un seul plus stable. De plus, on peut ainsi

évaluer la crédibilité de la réponse donnée à l'aide de la cohérence des résultats des différents arbres (ici, dans la pratique, une centaine).

3 Suivi de visage par recalage et ajustement de modèle 3D

Sachant que l'étape précédente renvoie l'information d'existence d'un visage dans la scène 3D analysée et celle d'un modèle générique 3D associé, la problématique considérée ici concerne la localisation, la reconstruction et le suivi du visage dans la scène 3D quelconque (d'intérieur ou d'extérieur), observée sous une direction inconnue et mobile, avec un éclairage quelconque. Le personnage est animé de mouvements 3D complexes et son visage, dans une attitude quelconque, peut représenter de 10 à 70 % de l'image.

Les principales méthodes proposées pour la reconnaissance du visage et le suivi de ses déformations dans une séquence d'images relèvent des approches markoviennes [17], de la morphologie mathématique [15], de l'apprentissage et la reconnaissance par réseaux de neurones [11], de formulations énergétiques par surfaces actives [28], par déformations de formes libres ou par superquadriques [3].

En l'absence de connaissance *a priori* sur les séquences à traiter, nous avons adopté ici une approche déterministe et géométrique. Celle-ci consiste à disposer d'un modèle générique 3D maillé de tête et à l'ajuster sur les images de la séquence. Nous avons développé une méthode d'initialisation intra-image générique et robuste à partir des techniques de la morphologie mathématique, puis une méthode de recalage fin 3D / 2D par extraction de primitives 2D d'image et 3D du modèle et optimisation sous contraintes. Nous présentons et discutons en terme de robustesse les résultats obtenus par cette approche sur les séquences Carphone, Foreman, Baby et Armel.

3.1 Ajustement 3D / 2D

Le modèle générique 3D de tête est représenté sous forme d'un maillage triangulé. Les yeux, la bouche et la base du cou sont des trous du maillage.

Le maillage est constitué de facettes planes (pas de triangles gauches). Une arête appartient soit à une seule facette (bord terminal), soit à deux facettes. Plusieurs facettes ne peuvent pas se raccorder en un point intérieur à une arête (*i.e.* pas de sommet en T). La topologie du maillage est quelconque : simplement connexe ou non, localement convexe ou non. A chaque facette, on associe la normale qui pointe vers l'extérieur de l'objet.

L'ajustement 3D/2D du modèle maillé sur l'image se fait à partir de primitives 2D et de primitives 3D. Les premières sont les discontinuités en niveaux de gris de l'image obtenues par un opérateur de Deriche seüllé avec hystérésis. Les secondes sont les contours occultants, com-

posés des bords terminaux (bords des trous formés par la bouche, les narines, les yeux, la base du cou) qui sont des caractéristiques intrinsèques à la surface, et les limbes qui sont des caractéristiques extrinsèques à la surface. Sur une surface G^1 continue, les limbes sont des lignes où la direction d'observation est tangente à la surface [20]. Sur un maillage, les limbes sont donc les arêtes communes à deux facettes adjacentes telles que les produits scalaires de leur normale avec la direction d'observation soient de signes contraires. En projection, les limbes et les bords terminaux forment un sur-ensemble de la silhouette : sur un visage observé de 3/4, le nez et les oreilles peuvent générer des limbes qui ne font pas partie de la silhouette. Dans le cas d'un objet non convexe, les limbes peuvent être cachés par d'autres parties de la surface. Nous proposons d'éliminer les limbes occultés par un algorithme de Z-buffer modifié. Par simplification, les arêtes partiellement vues sont considérées comme entièrement vues (figure 4).

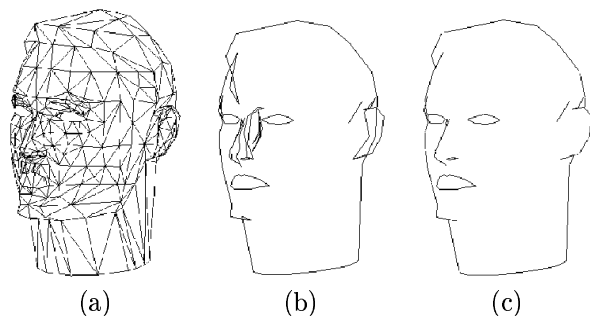


FIG. 4 – *Le modèle générique et ses contours occultants : (a) Modèle générique de tête : maillage triangulé (680 facettes). (b) Contours occultants. (c) Contours occultants vus après élimination des parties cachées par Z-buffer modifié.*

L'ajustement consiste à minimiser conjointement la distance entre les projections des contours occultants du modèle 3D et les contours. Nous minimisons la somme de ces distances sur les contours projetés. Pour accélérer le calcul de cette fonctionnelle, nous utilisons des cartes des distance \mathcal{D} [5] associées aux contours et marqueurs. Les paramètres de l'ajustement du modèle sont une mise à l'échelle, une translation 3D et une rotation 3D. Notons par q le vecteur de ces 7 paramètres, par $\vec{x}_I(q)$ la projection dans l'image d'un point du modèle et par $\mathcal{D}(\vec{x}_I(q))$ la valeur en ce point de la carte des distances.

L'ajustement est réalisé suivant la méthode itérative de Levenberg-Marquardt [21] [13]. La fonctionnelle à minimiser par rapport aux paramètres q est la somme des $\mathcal{D}(\vec{x}_I(q))$ pour les points des contours occultants du modèle (voir [8] pour les détails).

La nouvelle pose du modèle ainsi obtenue sert d'initialisation pour l'image suivante et le processus recommence. Cette approche conduit à des résultats très satisfaisants dès lors que l'initialisation est très proche de la solu-

tion finale (figure 9). Pour lever cet obstacle fondamental, nous avons développé une procédure d'initialisation intra-image générique et robuste que nous détaillons ci-dessous.

3.2 Procédure d'initialisation intra-image

Cette procédure fondée sur la morphologie mathématique comporte les étapes suivantes :

- partitionnement automatique de l'image 2D par ligne de partage des eaux sur le gradient filtré et étiquetage selon un critère de surface,
- localisation de la composante visage par critère de convexité et analyse des concavités à partir de l'enveloppe convexe et d'un squelette par zones d'influence combiné à un exosquelette,
- marquage des yeux et de la bouche qui sont des zones sombres localement contrastées à partir d'un opérateur de contraste généralisé et d'un coût de connexion [22],
- ajustement suivant des contraintes morphologiques (composantes du visage et marqueurs des yeux et de la bouche) puis de discontinuités (contours extraits par l'opérateur de Deriche).

L'ajustement est alors réalisé comme précédemment.

Les résultats obtenus à partir d'initialisations quelconques très éloignées de la solution finale démontrent l'excellente robustesse de la procédure aparamétrique proposée (figure 10).

4 Estimation de mouvement pour la reconstruction projective par facettes

Parallèlement à la tâche d'indexation/recalage, un processus de reconstruction projective est activé [4]. Nous nous sommes intéressés pour l'instant à des couples d'images extraites de la séquence que nous cherchons à segmenter en facettes planes. Ceci équivaut à segmenter au sens d'un mouvement homographique. Tout point d'une région vérifie alors le même modèle de mouvement. En raison de la complexité du modèle à estimer, nous avons divisé la segmentation en deux étapes. La première, que nous présentons ici, consiste à estimer un champ de disparité dense robuste et régularisé. La deuxième, en cours d'étude, vise à segmenter ce champ par estimations successives des différents modèles dominants.

Les techniques classiques basées sur des critères de corrélation [29] présentent des faiblesses pour résoudre notre problème (vecteurs disparité à composantes entières, pas de gestion des discontinuités, pas de régularisation du champ). Nous avons donc développé un algorithme dérivé de l'estimation du flot optique permettant d'obtenir

un champ estimé robuste et régularisé [16]. Par cet aspect, il rejoint certaines techniques déjà proposées [24] [7]. L'originalité de notre contribution réside dans l'utilisation de la géométrie épipolaire couplée à un schéma multirésolution mêlant méthodes différentielles et discrètes qui semblent a priori incompatibles. Le champ dense vérifiant la géométrie épipolaire, est géométriquement cohérent avec le modèle de projection perspective. L'utilisation d'un schéma multirésolution contraint permet d'assurer la convergence de l'algorithme pour un gain de temps important.

4.1 Principe général

Équation de base du flot optique. Soit $I_i(s)$ l'intensité lumineuse dans la i^{eme} image, où $s = (x, y)$ représente la position spatiale. Sous l'hypothèse de non variation de l'intensité d'un point le long d'une trajectoire, la DFD (Displaced Frame Difference) est égale à zéro :

$$DFD(s, ds) = I_1(s) - I_2(s + ds) = 0, \quad (1)$$

où $ds = (dx, dy)$ est le déplacement d'une image à l'autre d'un point matériel le long des axes x et y .

Telle quelle, cette équation ne prend pas en compte la géométrie de la prise de vue (importante pour une bonne reconstruction 3D). C'est pourquoi nous avons reformulé la DFD de façon à contraindre l'estimation par la géométrie épipolaire.

La géométrie épipolaire. Dans le cas d'un couple de vues d'une même scène statique, il existe une contrainte forte liant la projection m_1 d'un point de l'espace dans une image à une droite l_2 de correspondants potentiels dans l'autre image. Cette contrainte s'exprime sous forme matricielle en coordonnées homogènes :

$$\tilde{l}_2 = F_{12} \tilde{m}_1 \quad (2)$$

où F_{12} est une matrice 3×3 de rang 2. Cette matrice appelée matrice fondamentale [14] contient l'information géométrique maximale que l'on peut extraire d'un couple quelconque de vues non calibrées. La matrice fondamentale est déterminée à partir d'au moins huit paires de points en correspondance dans les deux images. En pratique, une extraction des points singuliers dans les deux images suivie d'une phase de mise en correspondance par corrélation est effectuée. La matrice F peut alors être calculée par diverses méthodes [14].

Nouvelle formulation de la DFD. En utilisant l'information épipolaire préalablement calculée, le vecteur disparité \vec{d}_s peut être décomposé sous la forme d'un vecteur normal \vec{N}_s et d'un vecteur tangent \vec{V}_s à la droite épipolaire associée à s (voir figure 5) : $\vec{d}_s = \vec{N}_s + \lambda_s \vec{V}_s$. On l'injecte alors dans l'équation (1) :

$$DFD(s, ds) = I_1(s) - I_2(s + \vec{N}_s + \lambda_s \vec{V}_s) = 0 \quad (3)$$

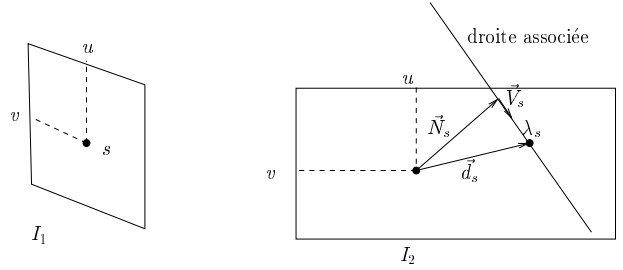


FIG. 5 – La contrainte épipolaire

Le vecteur normal \vec{N}_s et le vecteur tangent unitaire \vec{V}_s peuvent être calculés en tout point s en utilisant la matrice fondamentale. Le problème initial de recherche d'un champ de déplacement 2D est donc réduit à un problème 1D d'estimation d'abscisse λ_s le long des droites épipolaires.

4.2 Description de l'algorithme

Schéma multirésolution. La matrice fondamentale ne peut être correctement estimée que pour des déplacements importants entre deux prises de vues. À l'inverse, l'équation (3) est résolue en effectuant une linéarisation par développement limité par rapport à λ_s , ce qui suppose des déplacements faibles le long de la droite épipolaire. Afin de pouvoir coupler ces approches *a priori* incompatibles, un schéma multirésolution a été développé, l'estimation allant des basses résolutions (sommet de la pyramide) vers les hautes résolutions (base de la pyramide). À un niveau k de la pyramide, la disparité λ_s^k le long de la droite épipolaire est décomposée en une disparité λ_s^{k-1} , issue de la projection de l'estimation réalisée au niveau de plus basse résolution $k-1$ et d'un incrément $d\lambda_s^k$ à estimer. Le vecteur déplacement s'écrit alors : $\vec{d}_s^k = \vec{N}_s^k + \lambda_s^{k-1} \vec{V}_s^k + d\lambda_s^k \vec{V}_s^k$

L'équation (3) s'écrit alors pour un niveau k donné :

$$I_1^k(s) - I_2^k(s + \vec{N}_s^k + \lambda_s^{k-1} \vec{V}_s^k + d\lambda_s^k \vec{V}_s^k) = 0 \quad (4)$$

où l'inconnue est à présent $d\lambda_s^k$.

Les pyramides d'images I_1^k et I_2^k sont construites par filtrage puis sous-échantillonnage. Pour chaque niveau, la matrice fondamentale est calculée par un changement de base.

Méthode d'estimation régularisée. Nous nous plaçons maintenant à un niveau de résolution donné k . Pour des raisons de clarté l'indice k sera omis dans les équations qui suivent.

Afin d'estimer le champs de disparités, l'équation (4) est linéarisée vis à vis de l'inconnue $d\lambda_s$ autour de $s + \vec{N}_s + \lambda_s \vec{V}_s$. Le développement limité est maintenant valide puisque le vecteur incrément $d\lambda_s \vec{V}_s$ est faible devant le vecteur $\vec{N}_s + \lambda_s \vec{V}_s$.

$d\lambda_s$ est considéré comme une réalisation d'un champ de Markov aléatoire. Le meilleur champ de disparités en accord avec le critère Bayésien du *M.A.P.* (Maximum A

Posteriori) revient au problème de minimisation globale suivant :

$$\begin{aligned} \widehat{d\lambda} &= \arg \min_{d\lambda \in \mathbb{R}} H(d\lambda) \\ &= \arg \min_{d\lambda \in \mathbb{R}} (H_1(d\lambda) + \alpha[H_2(d\lambda)]) \end{aligned} \quad (5)$$

où α est une constante réelle ayant pour but d'équilibrer les deux termes énergétiques.

Le terme H_1 est le terme d'énergie lié aux observations. Il provient de la linéarisation de la DFD.

Le terme H_2 est le terme d'énergie qui vise à lisser localement le champ de vecteurs. H_2 favorise des vecteurs déplacements similaires d_s et d_r pour toute paire $\langle s, r \rangle$ de positions voisines.

Cependant la présence de zones d'occultations et de variations d'éclairément peut mettre en défaut l'hypothèse de conservation de la luminance. Parallèlement la présence de discontinuités de profondeur viole la contrainte de lissage. C'est pourquoi nous avons ajouté un estimateur robuste à chaque terme énergétique. Il nous permet alors de borner l'apport énergétique en cas de viol flagrant d'une hypothèse. Cela se traduit par une pondération de chaque terme : plus l'énergie est importante, plus le poids est faible voire nul si elle dépasse un seuil donné.

Résolution du problème de minimisation globale.

Un champ dense initial est calculé pour le niveau le plus haut de la pyramide. Celui-ci est obtenu par interpolation à partir des paires de points en correspondance nécessaires au calcul de la géométrie épipolaire.

À un niveau de résolution donné, un schéma itératif de Gauss-Seidel est mis en œuvre pour résoudre le problème de minimisation. Celle-ci est effectuée alternativement sur le champ de disparités $d\lambda_s$ et sur le champ des pondérations. La disparité en chaque pixel est successivement calculée en fixant les autres valeurs et en prenant le minimum local déduit de l'équation (5). On obtient ainsi une expression littérale de $d\lambda_s$.

5 Résultats

Nous présentons les résultats obtenus pour chacune des étapes décrites précédemment.

Pour la phase de reconnaissance, les résultats présentés s'appuient sur une des versions de la base de données Columbia contenant des images de 20 objets, chacun étant pris en photo sous 72 angles différents (de 5 degrés en 5 degrés). Les vues sont des images de définition 128×128 en 256 niveaux de gris. Le protocole consiste à utiliser la moitié des images (une sur deux) pour l'apprentissage et l'autre moitié pour le test (donc pour chacun des deux sous-ensembles nous avons des vues de 10 degrés en 10 degrés).

Dans un premier temps, l'algorithme a été testé dans les conditions restreintes de la base de Columbia seule. Nous avons utilisé 100 arbres, la tolérance angulaire pour les questions était de $\frac{\pi}{12}$, les ensembles aléatoires utilisés pendant l'apprentissage contenaient 200 questions. Nous

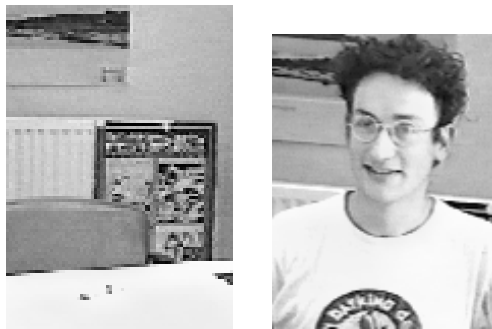


FIG. 6 – Deux des images ayant servi au test sur la séquence vidéo Armel. Le visage est détecté sans ambiguïté dans la seconde image, malgré la présence d'autres objets (T-shirt, fond).

Tags	Proportion d'images	Taux d'erreur
16 + 16	1 sur 2	0.41%
16 + 0	1 sur 2	1.1%
8 + 8	1 sur 2	1.25%
16 + 16	1 sur 4	3.98%
16 + 16	1 sur 8	8.1%

FIG. 7 – Taux d'erreur sur la base Columbia.

avons d'abord évalué l'importance du nombre d'images disponibles pendant l'apprentissage, et de la finesse de la quantification des angles dans la détection de bords.

Dans une première série d'expériences, nous avons utilisé la moitié des images pour l'apprentissage, en faisant varier le nombre de tags : soit 16 tags d'orientations, soit 16 tags d'orientations et 16 tags d'orientations grossières, et enfin 8 tags d'orientations et 8 tags d'orientations grossières. Dans une deuxième série, on a gardé le type de tags qui permettait d'obtenir le meilleur taux de reconnaissance, mais on a diminué fortement le nombre d'images utilisées pendant l'apprentissage (les tests étant toujours faits sur toutes les autres images de la base). Les taux d'erreurs, tout à fait satisfaisant, sont donnés fig. 7. La base d'images a ensuite été enrichie par 10 vues du sujet de la séquence Armel (voir Fig. 2).

Après reconstruction des arbres, nous avons testé l'indexeur sur trois ensembles d'images : (a) un ensemble de 20 imageries prises aléatoirement dans une partie de la séquence ne contenant pas le sujet, (b) un ensemble de 41 imageries contenant le visage du sujet dans la séquence, (c) les mêmes images que (b), mais sur lesquelles on a détourné grossièrement le sujet².

On voit que les images de fond (sans visage) produisent très peu de réponses positives. En contraste, les images "nettoyées" (série (c)) produisent un bon taux de détection, compte tenu de l'absence de contrôle fort sur les conditions de prise de vues. Les images brutes sont net-

2. cette segmentation approximative sera produite par la phase de détection des zones non-conformes dans les études futures.

Série	Taux de reconnaissance (classe "visage")
(a)	2%
(b)	51 %
(c)	77 %

FIG. 8 – Taux d’erreur sur les images extraites de la séquence analysée.

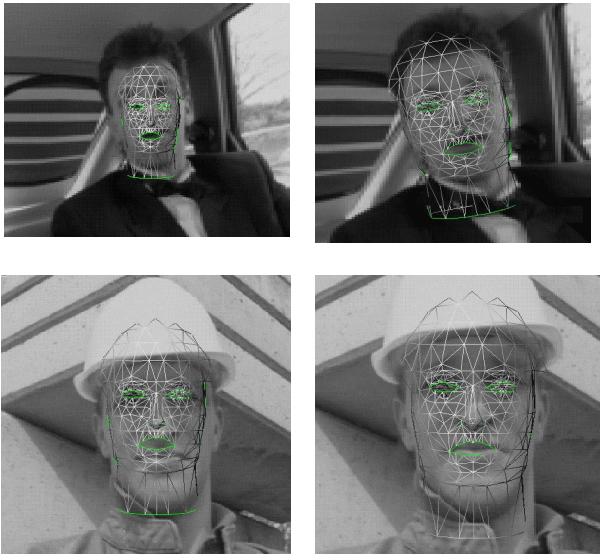


FIG. 9 – Exemples d’ajustement sur les séquences “carphone” (en haut) et “foreman” (en bas). Initialisation du modèle (à gauche) et ajustement réalisé (à droite, zoom).

tement moins performantes, ce que nous attribuons à un manque de robustesse du codage. On peut néanmoins supposer que dans une implémentation on-line, on enregistrerait la présence d’un visage si plus de 50% d’images consécutives répondent “oui” au test de présence.

Pour la tâche de recalage, la procédure d’initialisation intra-image a été mise en œuvre sur les séquences Carphone (89 images), Foreman (250 images), Baby (40 images) et Armel (100 images). Représentatives de différents environnements, de différentes conditions de prises de vue et offrant un large éventail de difficultés en termes de mouvements rapides, d’objets surgissant et d’occultations partielles, ces séquences de test montrent une bonne robustesse de la technique morphologique de segmentation du visage avec 3% d’erreur sur Carphone, 5% sur Foreman, 7% sur Baby et 15% sur Armel.

En ce qui concerne l’ajustement 3D/2D fin, qui dépend d’un paramètre pour le seuillage par hystérésis, réglé de façon ad hoc pour chaque séquence, il fournit une estimation précise des paramètres de pose du modèle comme l’établissent les tests sur des images de synthèse: erreur de 4% pour les translations et 12% pour les rotations. La tâche d’analyse du fond de la scène a été réalisée sur des images où le personnage n’était pas encore présent.

Sur le couple d’images présenté, le calcul de la géométrie épipolaire ainsi que l’initialisation par interpolation ont été réalisés à partir de 89 paires de points en correspondance extraits de façon automatique.

La figure (11) présente les résultats obtenus sur ces images. La carte de discontinuité montre les zones (points foncés) où l’énergie liée au terme de lissage a été bornée i.e. les discontinuités de profondeur. La carte liée aux observations révèle les zones où la DFD n’est pas vérifiée (points de trop fort gradient). Dans ces zones, le champ est lissé. On observe que la discontinuité de profondeur a bien été détectée sur le bord supérieur de la table, ainsi que la discontinuité d’orientation entre le mur droit et le mur du fond. Cependant leur localisation reste peu précise ce qui implique que l’étape de segmentation devra s’appuyer sur les contours dans l’image pour atteindre une localisation satisfaisante des discontinuités. Dans les zones homogènes, un champ de disparité cohérent est obtenu grâce à la régularisation introduite par le formalisme Markovien. L’influence du champ initial a été étudiée: elle est minorée par le processus multi-résolution qui permet des variations importantes entre les champs de disparité initial et final.

6 Conclusion et perspectives

En conclusion, notre contribution a porté sur :

- le développement d’une méthode d’indexation qui permet une reconnaissance d’un objet d’une base de données avec un taux d’erreur satisfaisant, comme le montrent les résultats sur la base Columbia. Lorsque la reconnaissance de visage est faite sur des images où peu d’éléments perturbateurs (fond texturé ou reste du corps) sont présents, la reconnaissance est également possible. L’analyse du mouvement devrait permettre de supprimer les zones conformes de façon satisfaisante. Il est aussi à noter qu’en l’absence de visage on n’obtient pratiquement jamais de fausse réponse positive. Nous concentrerons donc la suite de nos efforts à rendre le codage encore plus résistant aux perturbations, ainsi qu’à tester notre approche sur une base de visages plus vaste et dans des conditions expérimentales plus variées.
- le développement d’une méthode générique d’ajustement 3D/2D à partir d’une représentation maillée simplifiée de l’objet, intégrant une procédure d’initialisation intra-image robuste à partir de techniques de morphologie mathématique. Les prochaines contributions intégreront la notion de mouvement pour le suivi à l’ordre 1 de l’objet dans la séquence, la mise en œuvre d’un modèle de déformation et la prise en compte des informations de couleur.
- l’estimation dense conciliant approche différentielle et contraintes géométriques. Elle fournit un champ

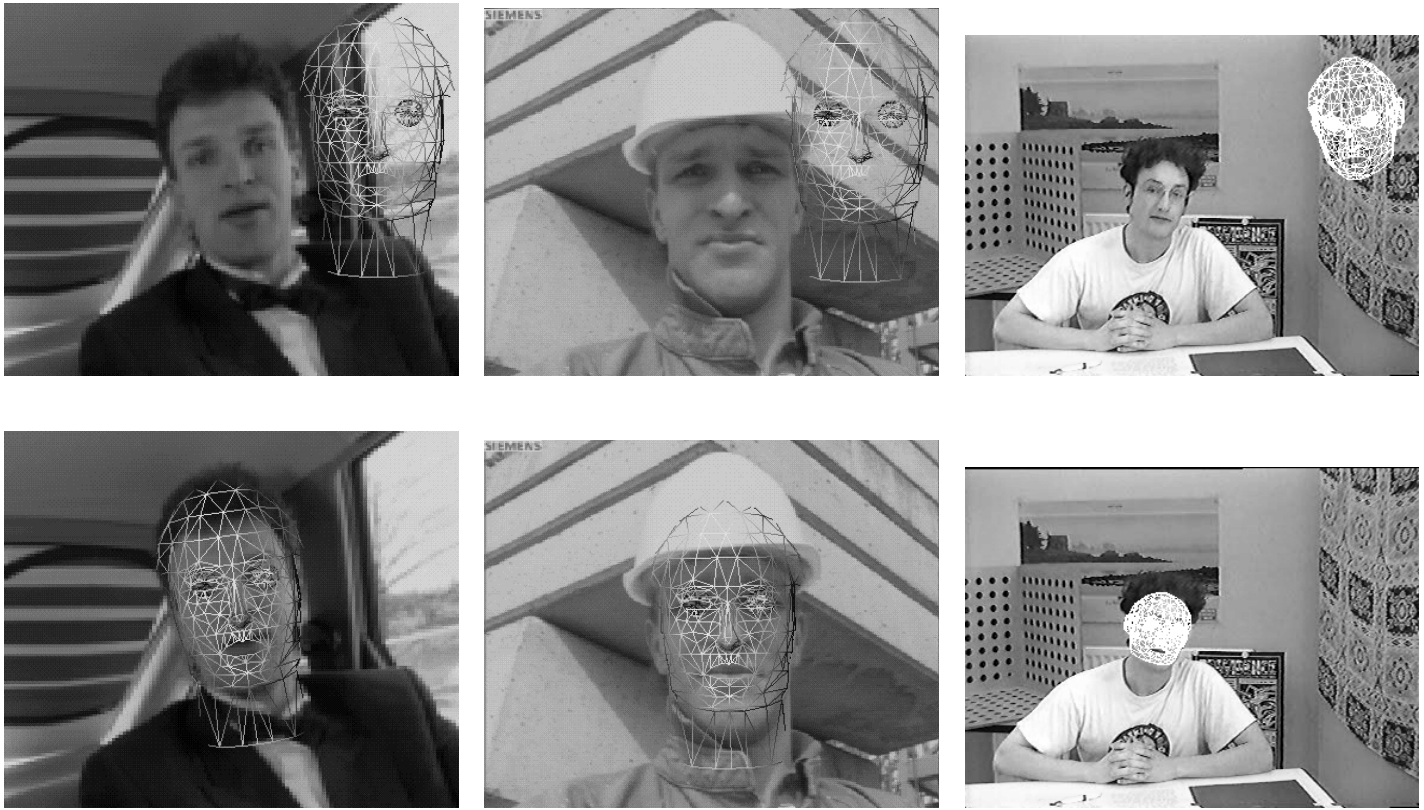


FIG. 10 – Exemples d'ajustement (ligne inférieure) à partir d'initialisation lointaine (ligne supérieure) sur les séquences Carphone, Foreman et Armel.

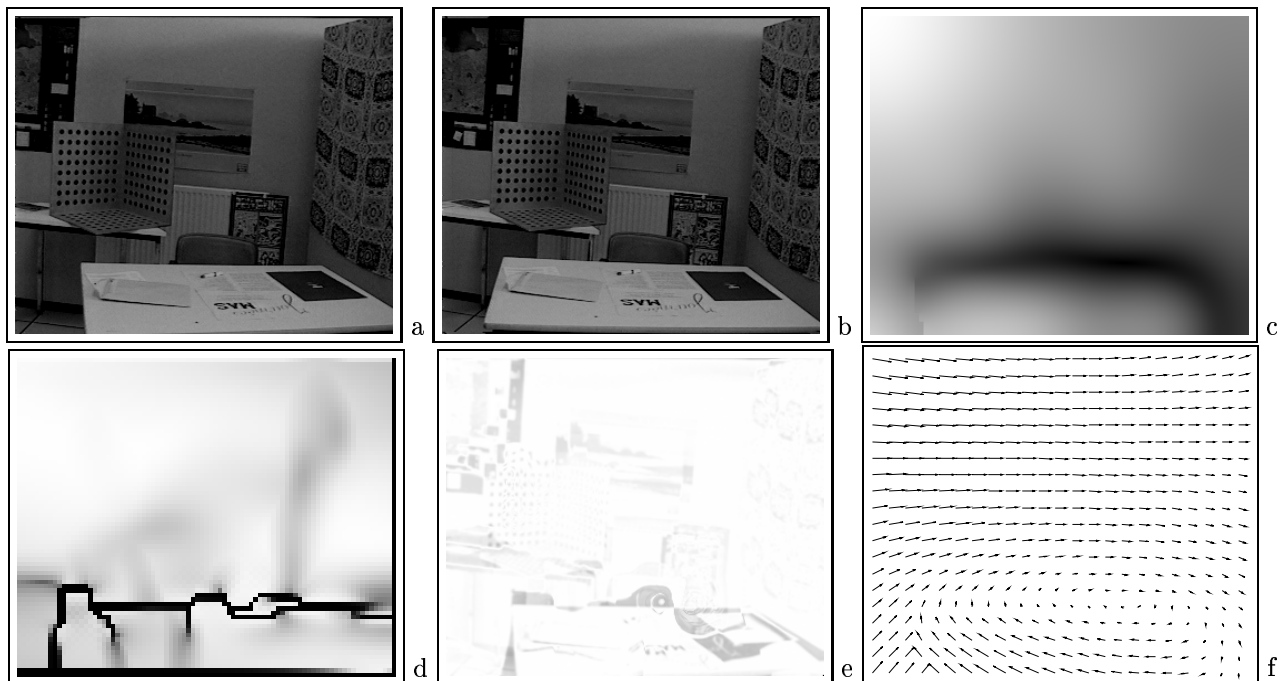


FIG. 11 – a,b : images originales gauche et droite - c : carte de disparité dense (plus un point est foncé, plus la disparité en ce point est importante) - d,e : cartes de discontinuité et d'observation - f : champ de disparité final

lissé à valeurs réelles autorisant cependant certaines discontinuités. De plus, ce champ est cohérent avec la géométrie épipolaire et donc avec un modèle de projection perspective, point important pour l'étape de reconstruction tridimensionnelle. La modélisation explicite des zones d'occlusion et leur prise en compte dans le critère d'énergie à minimiser pourrait encore améliorer les résultats obtenus.

De façon plus globale, au niveau de l'ensemble du projet, la restitution 3D d'une scène quelconque nécessite la réalisation des étapes de segmentation en facettes planes et de reconstruction projective des zones inconnues de la scène, ainsi que l'utilisation des objets connus pour le passage de la représentation dans un repère euclidien. Une généralisation de la méthode au cas de plusieurs objets connus présents dans la scène (correspondant au même modèle générique ou non) et fixes (ce qui implique une reconnaissance avec environnement complexe et un recalage de modèle 3D ne pouvant s'appuyer que sur les mouvements de la caméra) est actuellement à l'étude.

Références

- [1] Amit (Y.) et Geman (D.). – Shape recognition and quantization with randomized trees. *Neural Computation*, vol. 9, 1997, pp. 1545–1588.
- [2] Amit (Y.), Geman (D.) et Wilder (K.). – Recognizing shapes from simple queries about geometry. *PAMI*, 1997. – accepté pour publication.
- [3] Bardinet (E.). – Modèles déformables contraints : applications à l'imagerie cardiaque. – Thèse de doctorat - Univ. Paris IX - Dauphine, Déc. 1995.
- [4] Bobet (P.), Blanc (J.) et Mohr (R.). – Aspects cachés de la trilinearité. In : *RFIA*. – 1996.
- [5] Borgfors (G.). – Distance transformations in digital images. *CVGIP*, vol. 34, 1986, pp. 344–371.
- [6] Chiariglione (L.). – Mpeg-4: Coding of audio-visual objects. In : *ISO/IEC JTC1/SC29/WG11 N*, éd. par MPEG'96. ISO. – 1996.
- [7] Demarty (J.) et Schmitt (F.). – Reconstruction 3d dense à partir de séquences d'images. In : *Journées ORASIS 96*. – Clermont-Ferrand, 1996.
- [8] France-Telecom (édité par). – *CORESA '97*. CNET.
- [9] Garcia-garduno (V.). – Une approche de compression orientée-objets ... – Thèse de doctorat de l'université de Rennes-1, mai 1995.
- [10] Geman (D.) et Jedynek (B.). – An active testing model for tracking roads from satellite images. *PAMI*, vol. 18, 1996.
- [11] Iwata (H.) et Nagahashi (H.). – Motion tracking of color image sequences using neural networks. In : *VCIP97, San Jose*, pp. 200–210. – 1997.
- [12] Jedynek (B.) et Fleuret (F.). – Reconnaissance d'objets 3d à l'aide d'arbres de classification. In : *Proceedings of the 3rd international Conference IMAGE'COM*, pp. 25–30. – Bordeaux, 1996.
- [13] Lowe (D.). – Fitting parametrized three-dimensional models to images. *PAMI*, vol. 13, n° 5, Mai 1991, pp. 441–450.
- [14] Luong (Q.-T.) et Faugeras (O.). – The fundamental matrix: theory, algorithms, and stability analysis. *IJCV*, vol. 17, n° 1, 1995.
- [15] Marques (F.) et Molina (C.). – Object tracking for content-based functionalities. In : *VCIP97, San Jose*, pp. 190–199. – 1997.
- [16] Memin (E.) et Perez (P.). – Robust discontinuity-preserving model for estimating optical flow. In : *ICPR*. – Vienne, 1996.
- [17] Nefian (A.), Khosravi (M.) et Hayes (M.). – Real-time detection of human faces in uncontrolled environments. In : *VCIP97, San Jose*, pp. 211–219. – 1997.
- [18] Odobez (J.) et Bouthemy (P.). – Estimation robuste multi-échelle de modèles paramétrés de mouvement sur des scènes complexes. In : *RFIA*. – 1994.
- [19] Pereira (F.). – Mpeg-4: a new challenge for the representation of audio-visual information. In : *PCS96, Melbourne*. – 1996.
- [20] Ponce (J.) et Chelberg (D.). – Finding the limbs and cups of generalized cylinders. *IJCV*, vol. 1, 1987, pp. 195–210.
- [21] Press (W.), Teutolsky (S.), Vetterling (W.) et Flannery (B.). – *Numerical Recipes in C – Second Edition*. – Cambridge University Press, 1992.
- [22] Preteux (F.). – *On a distance function approach for gray-level mathematical morphology*, chap. 10. – Marcel Dekker Inc., 1991.
- [23] Quinlan (J.). – Induction of decision trees. *Machine Learning*, vol. 1, 1986.
- [24] Robert (L.) et Deriche (R.). – Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In : *ECCV'96*. – 1996.
- [25] Springer (édité par). – *Audio-and Video-based Biometric Person Authentication, Crans-Montana, Switzerland*. – mars 1997.
- [26] Suen (C.), Nadal (C.), Legault (R.), Mai (T.) et Lam (L.). – Computer recognition of unconstrained handwritten numerals. *Proc. IEEE*, vol. 80, 1992, pp. 1162–1180.
- [27] Swain (M.). – Object recognition from a large database using a decision tree. In : *Proc. of the DARPA Image Understanding Workshop, Vol II*, pp. 690–696. – Morgan Kaufman Publishers, 1988, avril 1988.
- [28] Terzopoulos (D.) et Metaxas (D.). – Dynamic 3d models with local and global deformations: deformable superquadrics. *PAMI*, vol. 13, n° 7, 1991.
- [29] Zhang (Z.). – *Le problème de la mise en correspondance : l'état de l'art*. – rapport de recherche n° 2146, INRIA, 1993.