Flot d'occupation 3D à partir de silhouettes latentes

Jean-Sébastien Franco
INRIA Sud-Ouest
LaBRI, France
franco@labri.fr

Li Guan
UNC Chapel Hill, USA
ETH-Zürich, Suisse
lguan@cs.unc.edu

Edmond Boyer
INRIA Rhône-Alpes
LJK, France
edmond.boyer@inrialpes.fr

Marc Pollefeys
ETH-Zürich, Suisse
UNC Chapel Hill, USA
marc@inf.ethz.ch

Résumé

Nous examinons l'extraction simultanée d'informations de forme et de mouvement, dans le contexte des systèmes multi-caméra calibrés. Nous proposons une nouvelle analyse basée sur l'information de silhouette latente dans les images. De nombreuses méthodes utilisent un modèle explicite de surface pour une telle analyse. Nous montrons qu'il est possible d'extraire une information pertinente sans modèle explicite, en utilisant une méthode EM pour simultanément extraire un ensemble de probabilités sur une grille de voxels représentant la scène et une estimation du champ de déplacements 3D entre deux pas de temps consécutifs. La méthode s'avère être un outil robuste pour des tâches d'inférences structurelles de plus haut niveau, comme l'extraction de parties en mouvement rigide, ou de squelette cinématique. Nous montrons expérimentalement l'utilité et la validité de la méthode.

Mots Clef

Flot 3D, Reconstruction 3D, Multi-vue, Silhouettes.

Abstract

In this paper we investigate shape and motion retrieval in the context of multi-camera systems and we propose a new low-level analysis based on latent silhouette cues. Many shape and motion analysis tools rely on the use of explicit surface models. Our analysis does not rely on explicit surface boundaries and uses an EM framework to simultaneously retrieve a set of volumetric voxel occupancy probabilities and a best estimate of the dense 3D motion field from the last consecutively observed multi-view frame set. As the framework uses only latent, probabilistic silhouette information, the method yields a promising 3D scene analysis tool robust to many sources of noise in difficult lighting and outdoor conditions. It can be used as input for higher level shape modeling and structural inference tasks. We demonstrate its practical use for shape and motion analysis experimentally.

Keywords

3D Flow, 3D Reconstruction, multi-view, silhouettes.

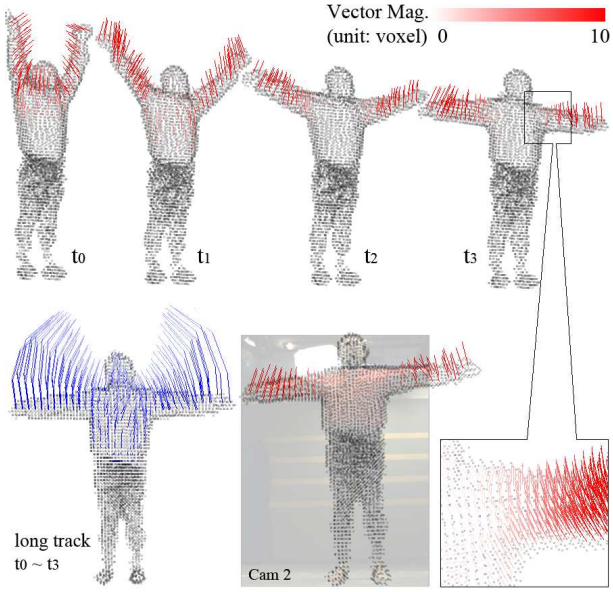


FIG. 1 – Flot 3D dense et grille d'occupation probabiliste 3D estimés. La grille a été seuillée à 98% pour visualiser l'objet sous-jacent estimé, seuls les vecteurs déplacement dans ce voisinage sont affichés (de norme codée en rouge). La trajectoire sur l'ensemble de la séquence de certains points est donnée en bleu. Une des 8 vues utilisées est superposée à la grille et aux déplacements du temps t_3.

1 Introduction

Nous proposons dans ce travail une nouvelle représentation et méthodologie pour l'analyse de mouvement de sujets dynamiques dans une scène observée par plusieurs caméras calibrées. Inspirée des techniques de flot optique 2D [18] et du travail initial sur le flot de scène de Vedula *et al.* [20], notre méthodologie permet l'estimation d'un flot de déplacement 3D entre deux jeux d'images consécutifs de la scène pour des objets arbitraires. Contrairement aux approches existantes cependant, le champ de déplacement que nous calculons est volumétrique et dense en 3D et ne repose pas sur une représentation explicite de la surface des objets de la scène. Seule une information latente des silhouettes dans les images capturées est utilisée, sans qu'il soit nécessaire de prendre une décision binaire

sur la segmentation de celle-ci. Une seule hypothèse de régularisation est utilisée, à savoir l'hypothèse de continuité spatiale du mouvement en 3D. La méthodologie explore donc quelles contraintes minimales peuvent être utilisées pour l'analyse de mouvement 3D. Les motivations sont multiples : permettre d'affiner l'estimation de forme dans le temps, accroître la robustesse d'estimation de traditionnelles techniques basées silhouette, en profitant de leurs avantages. Construire des représentations spatio-temporelles 4D des scènes observées est un challenge majeur en vision par ordinateur. De telles représentations sont souvent recherchées pour construire des représentations 3D cohérentes dans le temps et analyser le mouvement 3D dans les scènes observées. Les applications sont nombreuses : "Free-viewpoint video", acquisition automatique de modèles et de performances 3D, réalité virtuelle et interaction homme-machine, analyse de propriétés pour la reconnaissance et l'appariement de modèles 3D. Nous donnons ici un aperçu des différents travaux connexes.

1.1 Approches existantes

Le problème de construction de représentations géométriques 3D à partir de séquences vidéo a tout d'abord été traité séparément pour chaque pas de temps, en utilisant la photocohérence [2], l'appariement et la triangulation de points épars, ou les silhouettes [14]. Les méthodes à base de photocohérence et d'appariements peuvent être plus précises (reconstruction des cavités des objets) mais supposent généralement un bon contrôle de la luminosité de la scène, une résolution image importante, et apportent de l'information principalement pour les parties fortement texturées des surfaces observées. En outre elles nécessitent souvent un calibrage des couleurs des différentes caméras qui peut être fastidieux. Les méthodes de modélisation à partir de silhouettes sont devenues populaires pour l'acquisition 3D [15], en offrant une alternative généralement plus simple, plus rapide pour reconstruire les objets indépendamment de leur texture, au prix d'une précision plus faible, mais suffisante pour un large nombre d'applications.

Beaucoup d'approches de modélisation 3D à partir d'images utilisent des représentations surfaciques, qu'elles ajustes aux images. Mais des représentations alternatives ont émergé, comme les représentations volumiques. Celles-ci ont permis de traiter des problèmes de robustesse au bruit pour des données photométriques [2] ou silhouettes [8, 10], montrant une robustesse particulière en conditions difficiles en extérieur [2, 10]. Nous souhaitons tirer bénéficier de ces propriétés pour l'analyse de mouvement 3D basé silhouettes.

Approches surfaces. Capturer la dynamique 3D d'une scène et affiner des estimations de formes quelconques à partir de séquences d'images est un but difficile, nécessitant d'exploiter la cohérence temporelle de la scène. Dernièrement, des méthodes de suivi de maillage ont permis d'avancer les recherches, en proposant d'estimer des maillages géométriquement et temporellement cohérents sur une séquence multi-vue. Un grand nombre de ces méthodes ajustent des maillages fixes aux images pour le suivi [4]. Ces méthodes sont toutefois souvent particularisées pour le cas de formes spécifiques, tels que les humains [21], en faisant des hypothèses sur la géométrie ou la cinématique sous-jacente. D'autres méthodes estiment des surfaces plus générales et peuvent quelquefois traiter le problème de changement de topologie apparente dans la séquence [19]. Les données utilisées par l'ensemble de ces méthodes peuvent être le flot optique 2D [4] ou les points d'intérêt non denses [19]. D'autres méthodes utilisent les silhouettes pour construire des représentations spatio-temporelles volumiques de la scène dans le cas d'objets articulés [3].

Remarquablement, une grande majorité des méthodes existantes pour capturer la dynamique des objets utilise les silhouettes des objets estimés, s'appuyant sur le constat que le nombre de points d'intérêt pouvant être obtenus à partir de jeux de données usuels (vêtements de couleur homogène, peu de texture) est trop faible pour contraindre la solution. Certaines méthodes utilisent d'ailleurs les silhouettes seules, sans appariement ou photocohérence [21], illustrant leur pouvoir contraignant, que nous cherchons à exploiter ici en toute généralité. La majorité des méthodes citées n'ont également été testées qu'en milieu intérieur contrôlé. Nous pensons donc qu'une nouvelle technique d'analyse 3D bas niveau robuste, peut venir complémenter avantageusement les techniques existantes, et pour certaines peut servir de données d'entrée pour apporter des contraintes robustes supplémentaires.

Approches flot de scène. De toutes les techniques d'analyse de mouvement 3D, la méthode proposée a un lien plus particulier avec les approches de calcul de flot de scène. De telles méthodes produisent des vecteurs de déplacement, le plus souvent associés à une représentation de surface, incluant des voxels [20], surfaces implicites [16], cartes de disparité stéréo [22] ou maillages [4]. La majorité des approches de flot de scène suppose que la surface sous-jacente est déjà calculée [20, 4] ou construite simultanément [16]. Le travail originel sur le flot de scene [20] propose également une analyse dans le cas où la surface est inconnue, mais permet uniquement d'estimer une approximation du volume sans champ de mouvement associé.

En outre, les méthodes de flot de scène se basent sur l'estimation de dérivées spatiales dans les images, basées sur les différences finies, parfois déléguées à des méthodes de flot optique 2D existantes [20, 4]. Comme noté dans [17], ceci peut limiter ces approches à de petits déplacements, dans le domaine de validité des approximations. L'analyse que nous proposons repose uniquement sur l'information des silhouettes et ne suppose pas qu'une surface soit préalablement estimée, la rendant complémentaire des approches flot de scène existantes. Nous traitons explicitement le problème des déplacements larges dans notre approche.

1.2 Présentation de la méthode

Nous proposons une nouvelle méthode pour extraire des informations de forme et de mouvement sans a priori, représentation ou initialisation de surface, ou sur la structure du mouvement autre que sa continuité spatiale. En cela la technique proposée présente des similitudes algorithmiques avec les algorithmes de flot optique 2D [18]. La nature des solutions trouvées est similaire, avec un champ de mouvement 3D propagé continuement depuis les zones de mouvements observé, mais la méthode se distingue par le type de données traitées, ici exclusivement les silhouettes.

Nous utilisons une modélisation probabiliste sous forme de grille d'occupation [7] (§2). Les problème d'estimation de la forme et du mouvement de l'objet, trop difficiles à résoudre simultanément du fait d'un espace combiné d'états très grand, est décomposé à l'aide d'un algorithme EM (§3). Celui-ci alterne entre l'estimation de probabilités d'occupation pour chaque voxel dans le pas d'estimation (§3.2), et l'estimation d'un maximum a posteriori pour le champ de mouvement dans le pas de maximisation (§3.3). Nous montrons que ce dernier peut être réduit à un problème d'optimisation discret de MRF et résolu avec un algorithme multi-échelle (§4). Enfin nous validons la méthode avec plusieurs jeux de données synthétiques ou réels en milieu intérieur et extérieur (§5).

2 Formulation

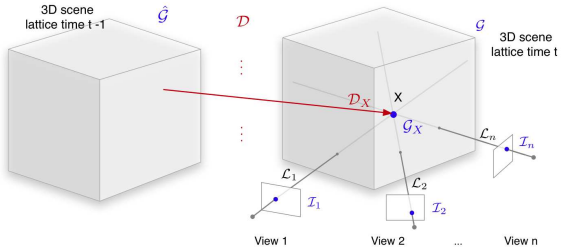


FIG. 2 – Variables statistiques et structure géométrique du problème. $\mathcal{G}_X$ est l'occupation au voxel X.

Nous représentons la scène avec une grille de points 3D (Fig. 2), notée $\mathcal{X}$. Au temps t, nous observons un jeu d'images $\mathcal{I}$, spécifiquement $\mathcal{I}_1, \mathcal{I}_2 \cdots \mathcal{I}_n$ de n vues calibrées (matrice de projection connue). Nous associons à chaque point $X \in \mathcal{X}$ un état d'occupation binaire, vide ou occupé, noté $\mathcal{G}_X \in \{0,1\}$. La conjonction de tous les états de la grille est noté $\mathcal{G}$. Nous souhaitons utiliser l'information du temps d'acquisition $t-1$, où l'état de la grille est noté $\hat{\mathcal{G}}$. Le mouvement de matière du temps $t-1$ à t est représenté par un champ de vecteurs $\mathcal{D}$. En particulier, nous associons à chaque point X le vecteur $\mathcal{D}_X$ qui déplace la matière du point $X-\mathcal{D}_X$ à X entre $t-1$ et t. Puisque la surface n'est pas explicitement représentée, nous supposons que ce champ de mouvement est défini partout dans l'espace et continu. Ceci revient à traiter l'espace comme un fluide contraint par les images, où le champ de mouvement déplace indifféremment la matière et l'air.

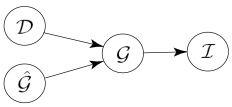


FIG. 3 – Dépendances entre groupes de variables du problème.

La relation entre les différentes variables du système peut être modélisée par la probabilité conjointe de ces variables : $p(\hat{\mathcal{G}}\mathcal{G}\mathcal{D}\mathcal{I})$. En se basant sur les dépendances de la Fig. 3, une décomposition de cette probabilité conjointe est donnée en (1) et modélise l'intuition : pour prédire les occupations $\mathcal{G}$, seules la connaissance des déplacements $\mathcal{D}$ et celle des occupations précédentes $\hat{\mathcal{G}}$ sont nécessaires ; pour prédire l'état des images $\mathcal{I}$, nous n'avons besoin que des occupations $\mathcal{G}$ au temps t. Nous faisons en outre l'hypothèse que les occupations $\mathcal{G}_X$ sont mutuellement indépendantes sachant $\mathcal{D}_X$ et $\hat{\mathcal{G}}_{X-\mathcal{D}_X}$. Ces dépendances et leur relation aux observations sont complètement analogues à l'interprétation probabiliste du modèle classique 2D de flot optique [18], où l'état de chaque pixel est prédit uniquement sachant le vecteur de flot optique et le pixel précédent correspondant.

$$p(\hat{\mathcal{G}}\mathcal{G}\mathcal{D}\mathcal{I}) = p(\mathcal{I}|\mathcal{G})p(\mathcal{G}|\mathcal{D},\hat{\mathcal{G}})p(\hat{\mathcal{G}})p(\mathcal{D}) \tag{1}$$

Nous faisons aussi l'hypothèse que les pixels sur lesquels un voxel de centre X se reprojette donnent des mesures indépendantes concernant ce voxel (pas de dépendances sur une ligne de vue), une simplification commune pour les méthodes à base de silhouettes. Nous supposons également que les mesures concernant un voxel sont conditionnellement indépendantes sachant l'état $\mathcal{G}_X$ du voxel X. Ainsi nous pouvons développer (1) :

$$p(\hat{\mathcal{G}}\mathcal{G}\mathcal{D}\mathcal{I}) = p(\mathcal{D})\prod_X \left(p(\hat{\mathcal{G}}_X)p(\mathcal{G}_X|\mathcal{D},\hat{\mathcal{G}}) \prod_i p(\mathcal{I}^i_{\mathsf{P}X}|\mathcal{G}_X) \right) \tag{2}$$

où i est l'indice de la vue. $\mathcal{I}^i_{\mathsf{P}X}$ est la couleur du pixel à la projection de X dans l'image i. Nous dénotons pour lisibilité le terme de mesure au temps t : $\mathbf{\Phi}(\mathcal{G}_X) = \prod_i p(\mathcal{I}^i_{\mathsf{P}X}|\mathcal{G}_X)$.
Dans (2), le terme $p(\hat{\mathcal{G}}_X)p(\mathcal{G}_X|\mathcal{D},\hat{\mathcal{G}})$ modélise l'information obtenue à l'instant précédent. En première approximation, seul un des déplacements, $\mathcal{D}_X$, influence X au temps t. De ce fait seul $\hat{\mathcal{G}}_{X-\mathcal{D}_X}$ influence $\mathcal{G}_X$:

$$p(\hat{\mathcal{G}}\mathcal{G}\mathcal{D}\mathcal{I}) = p(\mathcal{D})\prod_X \left(p(\hat{\mathcal{G}}_{X-\mathcal{D}_X})p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-\mathcal{D}_X}) \cdot \mathbf{\Phi}(\mathcal{G}_X) \right). \tag{3}$$

Le terme $p(\mathcal{D})$ modélise l'a priori sur le champ de déplacement 3D. Nous utiliserons la continuité du premier

ordre de celui-ci, comme décrit en §3.3. Nous supposons que l'information d'inférence du temps $t-1$ est disponible. Pour pouvoir utiliser celle-ci, nous traitons $\hat{\mathcal{G}}_X$ comme une variable latente. Ceci permet de conserver cette information probabiliste en marginalisant $\hat{\mathcal{G}}_X$ dans nos inférences.

3 Estimer le mouvement et l'occupation 3D

Pour résoudre le problème d'estimation, nous le traitons comme un problème d'estimation de $p(\mathcal{D}|\mathcal{I})$ à variables cachées $\mathcal{G}$, qui peut être résolu par un algorithme EM [6]. Bien qu'initialement utilisé pour des problèmes d'estimation de vraisemblance, il a été prouvé que l'EM [5] peut être appliqué pour résoudre des problèmes de maximum à posteriori (MAP-EM). Ceci permet d'ajouter au problème des a priori sur les variables à estimer, ce qui s'avère indispensable dans notre modélisation, pour incorporer la continuité spatiale du champ $\mathcal{D}$ dans notre problème. Dans notre cas, le MAP-EM consiste à construire une suite d'estimations $d^0, d^1 \cdots, d^*$ du champ de déplacement $\mathcal{D}$ (pas-M), tout en fournissant au $k+1$ème pas-E de convergence une estimation de $p(\mathcal{G}|\mathcal{I}, d^k)$. Ce dernier terme correspond aux occupations probabilistes des voxels de la grille, fournissant une représentation probabiliste de la matière analogue à d'autres méthodes [2, 8].

3.1 Développement de l'EM

Notre MAP-EM a pour but de trouver l'estimation d^* optimale de $\mathcal{D}$, telle que :

$$d^* = \text{argmax}_{\mathcal{D}}\ P(\mathcal{D}) \quad \text{avec} \quad P(\mathcal{D}) = p(\mathcal{D}|\mathcal{I}). \tag{4}$$

La solution est atteinte à partir d'une initialisation d^0, en construisant une suite $d^1, \cdots, d^*$ qui améliore la fonction objectif $P(\mathcal{D})$ (log-a-posteriori), i.e. $P(d^0) \leq \cdots \leq P(d^*)$. Dans la définition de l'EM [6] ce but est atteint en construisant une borne inférieure de $P(\mathcal{D})$, dont le maximum coïncide avec celui d'une fonction $Q(\mathcal{D}|d^k)$, analytiquement plus simple. L'estimation d^{k+1} est alors obtenue de la manière suivante :

$$\textbf{Pas-M :} \qquad d^{k+1} = \text{argmax}_{\mathcal{D}}\ Q(\mathcal{D}|d^k). \tag{5}$$

Pour le cas du MAP-EM, il a été démontré que la fonction $Q(\mathcal{D}|d^k)$ ayant ces propriétés prend la forme suivante [5] :

$$\begin{aligned} Q(\mathcal{D}|d^k) &= E_{\mathcal{G}|\mathcal{I},d^k}\{\ln p(\mathcal{I},\mathcal{G},\mathcal{D})\} \\ &= \sum_{\mathcal{G}} p(\mathcal{G}|\mathcal{I},d^k) \ln p(\mathcal{I},\mathcal{G},\mathcal{D}). \end{aligned} \tag{6}$$

Le **Pas-E** de l'algorithme consiste alors à évaluer le terme $p(\mathcal{G}|\mathcal{I}, d^k)$ de cette expression, i.e. les probabilités d'occupation sachant les données images et la précédente prédiction de $\mathcal{D}$. Nous calculons maintenant ces expressions à partir de la probabilité conjointe (3).

3.2 Pas-E : probabilités d'occupation

Exprimons les probabilités d'occupation de la grille $p(\mathcal{G}|\mathcal{I}, d^k)$ à partir de (3). La règle de Bayes est appliquée en (7) et l'expression factorisée en (8) en sortant les termes non dépendant des sommes. $\propto$ indique la proportionnalité à un facteur de normalisation unitaire près :

$$p(\mathcal{G}|\mathcal{I}, d^k) \propto \sum_{\hat{\mathcal{G}}} p(\hat{\mathcal{G}}\mathcal{G}d^k\mathcal{I}) \tag{7}$$

$$\propto \prod_X \mathbf{\Phi}(\mathcal{G}_X) \cdot \sum_{\hat{\mathcal{G}}_{X-d_X^k}} p(\hat{\mathcal{G}}_{X-d_X^k}) p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-d_X^k}), \tag{8}$$

où $\sum_{\hat{\mathcal{G}}_{X-d_X^k}} p(\hat{\mathcal{G}}_{X-d_X^k}) p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-d_X^k})$ marginalise les possibilités d'occupation du voxel $X-d_X^k$, le prédécesseur de X déplacé avec d_X^k. Nous affectons $p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-d_X^k})$ de manière déterministe : si le voxel prédécesseur $X - d_X^k$ était occupé (resp. vide), alors une fois déplacé en X il est toujours occupé (resp. vide) avec une probabilité 1. L'expression (8) devient :

$$p(\mathcal{G}|\mathcal{I}, d^k) \propto \prod_X \left(\mathbf{\Phi}(\mathcal{G}_X) \cdot p([\hat{\mathcal{G}}_{X-d_X^k} = \mathcal{G}_X]) \right). \tag{9}$$

Pour fournir une estimation de la forme 3D, la probabilité d'occupation de chaque voxel peut être identifiée après un pas-E en tant que $p(\mathcal{G}_X|\mathcal{I}, d^k) \propto \mathbf{\Phi}(\mathcal{G}_X) \cdot p([\hat{\mathcal{G}}_{X-d_X^k} = \mathcal{G}_X])$, le produit du terme d'observation au temps t, avec la probabilité du voxel prédécesseur de X.
$\mathbf{\Phi}(\mathcal{G}_X)$ peut être calculée en explicitant les termes de formation d'image $p(\mathcal{I}^i_{\mathsf{p}x}|\mathcal{G}_X)$. Pour chaque pixel x de chaque image, nous supposons avoir appris au préalable les paramètres $\mathcal{B}$ d'un modèle de fond quasi-statique de la scène sans objet d'intérêt. Le terme image $p(\mathcal{I}^i_{\mathsf{p}x}|\mathcal{G}_X)$ peut être affecté comme suit [8] :

$$p(\mathcal{I}^i_{\mathsf{p}x}|\mathcal{G}_X) = p(\mathcal{G}_X{=}0) p(\mathcal{I}^i_{\mathsf{p}x}|\mathcal{B}) + p(\mathcal{G}_X{=}1)\mathcal{U}(\mathcal{I}^i_{\mathsf{p}x}),$$

où $\mathcal{U}(\mathcal{I}^i_{\mathsf{p}x})$ est la distribution uniforme sur l'espace des couleurs, utilisée pour modéliser l'aspect des objets d'intérêt, dont nous ne connaissons pas l'apparence. $p(\mathcal{I}^i_{\mathsf{p}x}|\mathcal{B})$ est la probabilité pour $\mathcal{I}^i_{\mathsf{p}x}$ d'être tiré de la distribution du fond de paramètres $\mathcal{B}$. $p(\mathcal{I}^i_{\mathsf{p}x}|\mathcal{B})$ peut être, par exemple, Normale ou un mélange de Gaussiennes. L'information de silhouette est latente dans cette représentation, et ne requiert aucune décision de segmentation binaire.

3.3 Pas-M : champ de déplacement

Nous devons calculer l'expression de $Q(\mathcal{D}|d^k)$ dans l'équation (6). La distribution $p(\mathcal{I}, \mathcal{G}, \mathcal{D})$ peut-être calculée de manière similaire à (8), en factorisant la somme sur $\hat{\mathcal{G}}$ dans (3) :

$$p(\mathcal{I},\mathcal{G},\mathcal{D}) \propto p(\mathcal{D}) \prod_X \left(\mathbf{\Phi}(\mathcal{G}_X) \cdot p([\hat{\mathcal{G}}_{X-\mathcal{D}_X} = \mathcal{G}_X]) \right) \tag{10}$$

En prenant le logarithme de l'expression précédente pour le calcul de $Q(\mathcal{D}|d^k)$, et en notant que le terme $\mathbf{\Phi}(\mathcal{G}_X)$ ne dépend pas de $\mathcal{D}$, le pas-M devient :

$$\begin{aligned} d^{k+1} = &\operatorname{argmax}_{\mathcal{D}} \ \ln(p(\mathcal{D})) \\ &+ \sum_X \sum_{\mathcal{G}_X} p(\mathcal{G}_X|\mathcal{I},d^k) \cdot \ln p([\hat{\mathcal{G}}_{X-\mathcal{D}_X} = \mathcal{G}_X]) \end{aligned} \tag{11}$$

où $p(\mathcal{G}_X|\mathcal{I},d^k)$ est calculée dans le pas-E (équation (9)). Pour modéliser la continuité spatiale du champ, nous considérons pour l'a priori $p(\mathcal{D})$, que $\mathcal{D}$ est un champ de markov (MRF). De t à $t+1$, nous discrétisons le déplacement $\mathcal{D}_X$ en chaque point X à un ensemble de n possibilités $\mathcal{L} = \{l^1, \cdots, l^n\}$, ramenant (11) à un problème d'optimisation de graphe avec l'énergie suivante :

$$E_{MRF} = \sum_X \sum_{Y \in \mathcal{N}(X)} E_{XY}(l_X, l_Y) + \sum_X E_X(l_X), \tag{12}$$

où $\mathcal{N}(X)$ est le voisinage du point X dans le graphe, et où les termes binaires $\sum_X \sum_{Y\in\mathcal{N}(X)} E_{XY}(l_X,l_Y)$ et unaires $\sum_X E_X(l_X)$ peuvent être identifiés aux termes $\ln(p(\mathcal{D}))$ et $\sum_X \sum_{\mathcal{G}_X} p(\mathcal{G}|\mathcal{I},d^k) \cdot \ln p([\hat{\mathcal{G}}_{X-\mathcal{D}_X} = \mathcal{G}_X])$ respectivement dans (11).
Le pas-M de notre EM devient un problème d'étiquetage discret, avec le but de calculer un étiquetage $L \in \mathcal{L}^{|\mathcal{X}|}$, qui affecte à chaque noeud $X \in \mathcal{X}$ de la grille une étiquette dans $\mathcal{L}$ qui maximise l'énergie E_{MRF}, ce que l'on note :

$$d^{k+1} = \operatorname{argmin}_L E_{MRF}. \tag{13}$$

La solution de ce MRF donne la mise à jour du champ de déplacements. Nous détaillons ci-dessous l'implémentation et les propriétés de cette optimisation.

4 Optimisation du champ de déplacement

Du fait du large espace d'états du problème, et de la possibilité pour cet EM de rester bloqué dans des minimas locaux, des mesures additionnelles doivent être prises pour assurer la convergence de l'algorithme. Nous décrivons en détail les hypothèses faites sur le champ de déplacement et leurs implications sur les calculs (§4.1). L'algorithme « Fast-PD » [13] est indiqué pour l'optimisation (§4.2). Nous appliquons Fast-PD avec un schéma multi-échelle pour la stabilité, l'efficacité et la convergence de la méthode (§4.3).

4.1 Propriétés du champ

Nous supposons seulement que les déplacements sont bornés et le champ continu. Comme nous ne représentons pas la surface des objets et stockons uniquement une information probabiliste, notre champ de déplacement est volumique et couvre toute la grille. Donc la matière et l'air sont indifféremment englobés dans ce champ, tout en étant contraint par les images à déplacer des voxels probablement occupés à $t-1$ (resp. vide) vers des voxels probablement occupés (resp. vide) à t, ce qui est modélisé en (9). Il est possible de définir une continuité simple dans le terme binaire d'énergie V_{XY} dans (12), comme une distance calculant la norme de différences de vecteurs [9] :

$$E_{XY}(l_X,l_Y) = \lambda_{XY} \left| \boldsymbol{d}(l_X) - \boldsymbol{d}(l_Y) \right|^{0.8} \tag{14}$$

où λ_{XY} est un paramètre de poids, $\boldsymbol{d}(l)$ est le vecteur de déplacement représenté par l'état l, et le coefficient 0.8 est spécifiquement choisit en dessous de 1. Ce dernier est motivé par les statistiques de différences de vélocité étudiées pour les contraintes de flot optique [18].
Toutefois, de meilleures propriétés peuvent être obtenues pour éviter les déformations trop semblables à celles d'un fluide, dans le cas d'approches itératives multi-échelles [9], avec le terme binaire suivant :

$$E_{XY}(l_X,l_Y) = \lambda_{XY} \left| \boldsymbol{D}_X + \boldsymbol{d}(l_X) - \boldsymbol{D}_Y - \boldsymbol{d}(l_Y) \right|^{0.8} \tag{15}$$

où $\boldsymbol{D}_X$ et $\boldsymbol{D}_Y$ sont les déplacements estimés en positions X et Y à l'itération précédente.

4.2 Optimisation Fast-PD

Etant donnée la forme de nos termes d'énergie, la minimisation de (12) dans le pas-M peut être résolue avec des algorithmes d'optimisation discrètes de graphes. Nous choisissons l'approche Fast-PD [13], qui est particulièrement adaptée pour trouver des solution quasi-optimales pour une large classe de MRF NP-durs [12]. Cette approche a plusieurs avantages : elle est plus rapide que les algorithmes à coupure de graphe avec α-expansion [1] utilisés jusqu'alors, et garantit une borne supérieure sur la différence d'énergie entre la solution trouvée et la solution optimale. De plus elle permet de traiter des termes binaires d'énergie arbitraires, en levant la contrainte de sous-modularité des approches précédentes [11]. C'est ce qui nous permet de donner des formes plus élaborées pour le terme de continuité de champ de déplacement $E_{XY}(l_X,l_Y)$, tel que celui proposé en (15).

4.3 Approche multi-échelles

Pour éviter les minima locaux, accélerer la méthode et casser le problème en sous-problèmes moins gourmands en mémoire, nous optons pour une approche multi-échelles. Nous initialisons l'EM avec un premier alignement grossier en translation. Nous utilisons une paramétrisation

hiérarchique du volume, utilisée pour les problèmes d'appariement volumétriques en imagerie médicale [9]. La paramétrisation utilise une déformation de forme libre (FFD) à base de B-splines 3D, déformée par une grille de points de contrôle à résolution plus grossière que la grille initiale. A chaque échelle choisie, le MRF précédemment défini est résolu à la résolution correspondante des points de contrôle, et l'initialisation des échelles plus fines obtenue par interpolation du résultat plus grossier. L'espacement d_s des points de contrôle à chaque échelle est défini relativement à l'échelle globale de la scène. Nous contraignons les déplacements possibles pour un point de contrôle à être à une distance maximale de $d_s/2$. Ceci permet d'éviter les repliements du volume et garantit que la transformation volumique est un difféomorphisme. Plus spécifiquement, nous échantillonnons le volume cubique sur $[-d_s/2, d_s/2]$ autour d'un point de contrôle sur les trois axes. Par ailleurs nous répétons l'optimisation de la déformation à chaque échelle jusqu'à ce que le résultat obtenu du Fast-PD soit inchangé (en pratique 4 à 8 fois selon le jeu de données utilisé) pour une meilleur stabilité. Cela permet également de retrouver des champs de déplacements plus larges que l'intervalle de contrainte de l'échelle la plus grossière, ce qui est illustré dans nos résultats.

5 Résultats

Nous avons testé l'algorithme proposé sur des données réelles et synthétiques (voir les vidéos complémentaires du papier[1]). Ces données sont difficiles pour les algorithmes de l'état de l'art en estimation de mouvement 3D : les surfaces sont peu texturées, il y a du bruit de mesure et d'éclairage, et aucune calibration photométrique entre les différentes vues n'est effectuée.

Dans toutes les expériences réalisées, nous utilisons une grille d'occupation de résolution 128^3 et trois niveaux de grilles de contrôle pour la stratégie multi-échelles. Les plus grand déplacements admis sont de 5, 3, et 1 voxels respectivement pour les échelles choisies. Les points de contrôles sont donc espacés de 11, 7 et 1 voxels respectivement. L'EM converge en moins de 3 itérations pour tous les jeux de données présentés. Chaque pas de maximisation est répété de 4 à 8 fois par échelle. Le temps de calcul pour notre implémentation préliminaire est de plusieurs minutes par pas de temps pour la plupart des jeu de données. Plus de 9/10ème du temps est passé à construire les graphes d'optimisation et ses poids et non dans l'optimisation elle-même, suggérant une grande possibilité d'optimisation du code si celui-ci était spécialisé pour ce problème.

5.1 Données synthétiques

Deux cubes de taille différente se déplacent sur une trajectoire elliptique. Ils sont observés par 9 caméras virtuelles entourant la scène (Fig. 4(b)). Nous calculons la grille d'occupation probabiliste pour 9 pas de temps, à partir des images synthétiques en ajoutant un bruit gaussien sur les couleurs. 8 champs de déplacements sont calculés avec la méthode proposée. Pour illustrer la cohérence de la méthode, nous traçons en bleu des points du premier instant dont la probabilité d'occupation est supérieure à 0.98 dans la Fig. 4(a). Les traces des points obtenues avec les véritables déplacements sont également données en noir.

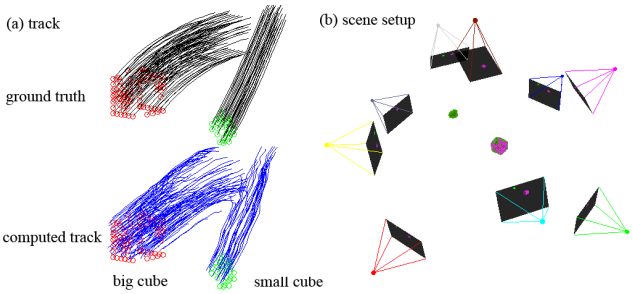


FIG. 4 – A gauche : les traces de déplacement calculées pour 2 cubes en mouvement. La vérité terrain est illustrée au-dessus. A droite : scène 3D utilisée.

Nous illustrons la justesse de la méthode en calculant l'erreur angulaire absolue (AAE) entre les véritables déplacements et ceux estimés pour les points de probabilité initiale d'occupation supérieure à 0.98 (donc très probablement dans l'objet). Nous montrons dans la Fig. 5 les statistiques obtenues pour chaque cube à chaque instant. L'AAE du petit cube est beaucoup plus faible que pour le grand cube. Nous interprétons ces résultats comme suit : (1) contrairement au méthodes basées sur la stéréo, la méthode proposée n'a pas accès aux composantes tangentielles de mouvement des surfaces observées. De ce fait, les rotations des objets sont plus difficiles à estimer par notre méthode, surtout lorsqu'elles sont autour d'axes propres de l'objet. Ceci peut être observé dans la figure Fig. 4, où les traces de mouvement du grand cube sont plus courtes et moins arquées que la vérité terrain, proposant une solution de translation localement valide. (2) L'occupation de voxels correspondant à de grands objets sont estimés davantage par interpolation que par une information image directe. Réciproquement, le mouvement du petit cube et la quantité d'information apportée par les silhouettes sont importantes par rapport à sa taille, réduisant ainsi l'ambiguïté de mouvement pour ces voxels. Ceci suggère que la méthode fonctionne mieux lorsque les silhouettes d'une partie fine d'un objet est observée en mouvement dans plusieurs images. Les mouvements reconstructibles par la méthode incluent donc les translations et les rotations observables depuis les silhouettes, soit la plupart des rotations articulaires, ce que nous illustrons dans la Fig. 1.

5.2 Données acquises en intérieur

Nous traitons la séquence utilisée par [8], qui inclut des mouvements de bras et de marche, respectivement analysés en Fig. 1 et Fig. 6. Il s'agit de 8 vidéos capturées à 15Hz. Due à une fréquence d'acquisition basse, les mouvements peuvent être relativement importants (plus de 5 voxels).

[1] www.labri.fr/~franco

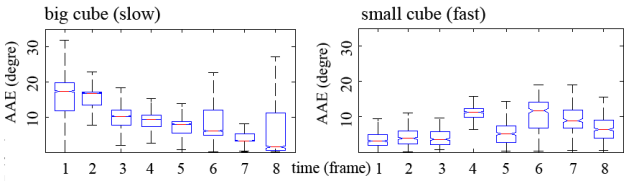


FIG. 5 – Erreur angulaire absolue (AAE) entre les vecteurs du champ de déplacement calculé et réel. Ce statistiques sont calculées sur les points de probabilité d'occupation 0.98. Le champ de vitesse du cube plus petit et rapide est mieux estimé.

Néanmoins notre méthode multi-échelles retrouve les large déplacements correctement. Les traces de quelques voxels sont fournies dans les figures en suivant les indications du calcul de champ, effectué sur chaque paire d'instants dans la séquence. L'apparente linéarité par morceaux des traces n'est pas un artefact de la méthode, mais montre les pas de mouvements réels entre deux instants de la séquence du fait de la basse fréquence d'acquisition.

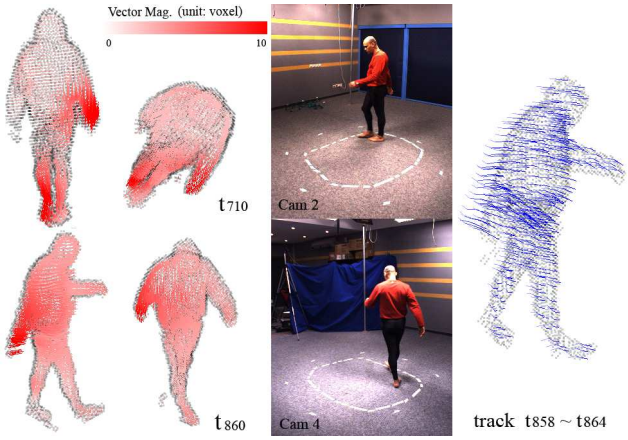


FIG. 6 – A gauche : 2 vues du champ de déplacement avec la grille de probabilité (> 0.98) seuillée en surimpression. La norme du déplacement est codée en rouge. Le mouvement plus large est obtenu au cours du balancement des membres pendant la marche. Milieu : images sources correspondantes. A droite : traces accumulées de t_{858} à t_{864}, correspondant au mouvement le plus large de la séquence (bras droit).

Fig. 7 montre 2 tranches de la grille d'occupation pendant le mouvement de bras. Les probabilités plus grandes sont plus foncées. Le champ de déplacement est en surimpression. Ceci montre la nature dense des données estimées. Cette représentation se montre robuste au bruit et aux occultations, comme le montrent la Fig. 1 où un bras sort du champ de vision sans pour autant être mal estimé par ailleurs. Un autre résultat, sur la séquence BABY, illustre les avantages de ne pas dépendre d'un modèle surfacique a priori. Bien que n'ayant accès à une vérification par la vérité terrain, les déplacements et leur amplitude estimés sont qualitativement cohérents.

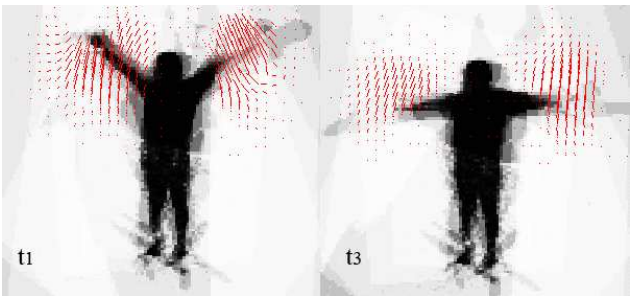


FIG. 7 – A gauche : mêmes tranches d'occupation à deux instants t_1 and t_3. Le champ est calculé sur le volume entier. Les mains descendent au cours de cet intervalle. Les vues 3D correspondantes sont dans la Fig. 1.

5.3 Données en extérieur

Nous traitons plusieurs jeux de données utilisé dans [10], comportant 6 à 9 vidéos à $30Hz$. Les couleurs ne sont pas calibrées, et les séquences comportent des difficultés supplémentaires : changement de luminosité, ombres, réflexions sur la sculpture métallique. Dû au bruit et au fond statique initialement acquis, des faux positifs de la forme apparaissent du fait des ombres au sol (Fig. 8(a)). Néanmoins la méthode estime une forme et un mouvement plausibles. Une méthode utilisant un modèle explicite de surface serait en difficulté dans cette situation en cherchant à englober le volume d'ombre. Les résultats de la séquence BENCH, illustrent le comportement de la méthode pour le cas où plusieurs personnes sont observées en milieu extérieur. Ils confirment qualitativement le bon comportement de la méthodes aux occultations et conditions d'acquisition difficiles.

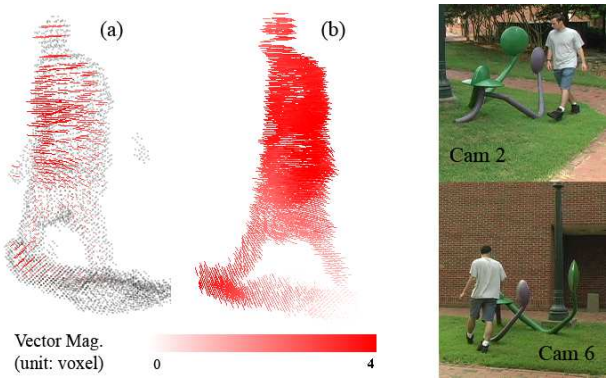


FIG. 8 – A gauche : occupation estimée à t_{259} et champ de déplacement entre t_{258} et t_{259}. (a) montre la direction et (b) la norme. A droite : 2 vues à t_{259}. L'échelle du codage couleur utilisé est différente de ROND du fait de l'acquisition 2 fois plus rapide ici.

La Fig. 9 montre le bénéfice potentiel de l'estimation conjointe de forme et de mouvement. En supposant dis-

poser de silhouettes parfaites à t_{258} (segmentées manuellement pour ce test), nous appliquons la méthode entre t_{258} et t_{259}. L'information correcte de t_{258} est visiblement propagée à t_{259}, faisant disparaître les artefacts et faux positifs du volume initialement dû aux ombres. Les occupations de t_{258} donnent un a priori d'occupation propagé pour les pas de temps suivants. Ceci suggère la possibilité de suivi et de raffinement de l'estimation d'une seule et même grille d'occupation et ses déplacements intermédiaires au cours du temps, pour de futurs travaux.

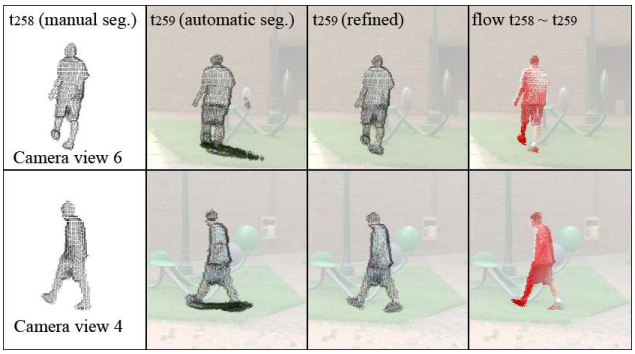


FIG. 9 – Application au raffinement d'occupation. Col. 1 : grille t_{258} calculées à partir de silhouettes manuellement segmentée. Col. 2 : occupation estimée par notre méthode à t_{259}. Les faux positifs dûs aux ombres et réflexions sont apparents. Occupation (col. 3) et champ de déplacement à t_{259} (col. 4) en utilisant la segmentation parfaite à t_{258}, dessinés pour les voxels de probabilité > 0.98. Les vues de t_{259} sont montrées en surimpression.

6 Discussion

Nous avons exploré une nouvelle direction pour l'analyse de forme et de mouvement 3D. Nous proposons une nouvelle approche pour estimer simultanément un champ de déplacement 3D dense entre deux instants et une estimation probabiliste de la forme 3D observée. L'approche n'utilise que des informations de silhouette latente, et un simple a priori de continuité spatiale, et se destine donc à des scènes et conditions très générales. Nos expérimentations montrent la validité de l'approche pour des données synthétiques et réelles en intérieur et extérieur avec des conditions difficiles. La méthode semble prometteuse pour de nouvelles possibilités et applications d'analyse de scène : segmentation de mouvement, suivi 3D, inférence de structure cinématique, estimation et raffinement de formes 3D au cours du temps. Les méthodes de modélisation et suivi existantes peuvent certainement utiliser nos champs et estimations en entrée, pour remplacer ou complémenter d'autres types de constructions bas-niveau comme le flot optique 2D ou les appariements non denses de points d'intérêt, en bénéficiant des avantages propres aux silhouettes. Davantage de données observées pourraient être inclues (profondeurs de Z-cameras ou stéréo par exemple) dans la méthodologie de fusion Bayésienne proposée. Celle-ci pourrait également être étendue pour affiner l'estimation d'une même forme au cours du temps, en utilisant des a priori de forme ou davantage d'observations passées.

Références

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *PAMI*, 2001.

[2] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the Space Carving algorithm. In *ICCV*, volume I, pages 388–393, 2001.

[3] G. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *CVPR*, June 2003.

[4] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *CVPR*, pages 1–8, 2007.

[5] F. Dellaert. The expectation maximization algorithm. Technical report, College of Computing, Georgia Institute of Technology, 2002.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorith. In *Journal of the Royal Statistical Society : Series B*, 1977.

[7] A. Elfes. Using occupancy grids for mobile robot perception and navigation. *IEEE Computer, Special Issue on Autonomous Intelligent Machines*, 22(6) :46–57, June 1989.

[8] J.-S. Franco and E. Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. In *ICCV*, volume 2, pages 1747–1753, oct 2005.

[9] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through mrfs and efficient linear programming. In *Medical Image Analysis*, 2008.

[10] L. Guan, J.-S. Franco, and M. Pollefeys. 3d occlusion inference from silhouette cues. In *CVPR*. IEEE Computer Society, 2007.

[11] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts ? In *PAMI*, 2004.

[12] N. Komodakis and G. Tziritas. Approximate labeling via graph-cuts based on linear programming. In *PAMI*, 2007.

[13] N. Komodakis, G. Tziritas, and N. Paragios. Fast, approximately optimal solutions for single and dynamic mrfs. In *CVPR*, June 2007.

[14] A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *PAMI*, 16(2) :150–162, Feb. 1994.

[15] S. Lazebnik, Y. Furukawa, and J. Ponce. Projective visual hulls. *IJCV*, 74(2) :137–165, 2007.

[16] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *IJCV*, 72(2) :179–193, 2007.

[17] J. Starck and A. Hilton. Correspondence labelling for wide-timeframe free-form surface matching. In *ICCV*, 2007.

[18] D. Sun, S. Roth, J. Lewis, and M. Black. Learning optical flow. In *ECCV*, 2008.

[19] K. Varanasi, A. Zaharescu, E. Boyer, and R. Horaud. Temporal surface tracking using mesh evolution. In *ECCV*, 2008.

[20] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *PAMI*, 27(3) :475–480, 2005.

[21] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. Articulated mesh animation from multi-view silhouettes. In *SIGGRAPH*, 2008.

[22] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, and D. Cremers. Efficient dense 3d scene flow from sparse or dense stereo data. In *ECCV*, oct 2008.