Reconstruction 3D de surfaces déformables par un modèle de faible rang hiérarchique

V. Gay-Bellile[1,2] A. Bartoli[1] U. Castellani[3] J. Peyras[1] S. Olsen[4] P. Sayd[2]

[1] LASMEA, UMR 6602 CNRS/UBP, Clermont-Ferrand

[2] CEA, LIST, Boîte Courrier 94, Gif-sur-Yvette, F-91191 France ;

[3] VIPS, Verona

[4] DIKU, Copenhagen

Vincent.Gay-Bellile@cea.fr

Résumé

Dans cet article, nous abordons le problème de la reconstruction 3D de surfaces déformables à partir de points mis en correspondance sur plusieurs images. Nous utilisons le modèle de faible rang : les déformations sont approximées par une combinaison linéaire de modes de déformation. Une nouvelle manière de représenter ce modèle est proposée : les modes de déformation sont ordonnés en fonction de l'amplitude des déformations qu'ils capturent. Cette représentation présente de nombreux avantages. Elle permet de lever certaines ambiguïtés et introduit un algorithme d'estimation hiérarchique du modèle, facilitant notamment la sélection automatique du nombre de modes par validation-croisée. Des résultats expérimentaux sur des vidéos variées montrent que l'approche proposée reconstruit des déformations plausibles de la surface observée, permettant l'incrustation réaliste d'objets virtuels 3D sur une vidéo d'une surface déformable.

Mots clefs

Reconstruction tridimensionnelle, modèle de faible rang, réalité augmentée.

1 Introduction

L'estimation de la structure et du mouvement à partir d'une vidéo prise par une seule caméra est un problème largement étudié en vision par ordinateur. Pour une scène statique (rigide), les rayons de vue associés au même point 3D observé par la caméra à des positions différentes s'intersectent dans l'espace. Cette propriété permet de définir des contraintes fortes sur la reconstruction. La reconstruction 3D de scènes rigides est en général un problème bien posé. Pour une scène dynamique, l'hypothèse que les rayons de vue s'intersectent n'est plus valide : évaluer la surface 3D, ses déformations et le mouvement de la caméra à partir d'une vidéo est dans la plupart des cas un problème sous contraint. Afin de garantir la reconstruction de déformations plausibles, il est alors indispensable de limiter l'espace des déformations admissibles. Une possibilité fréquemment utilisée dans la littérature consiste à approximer les déformations d'une surface par une combinaison linéaire de modes de déformation. Les algorithmes de reconstruction utilisant le modèle de faible rang, par exemple [1, 3, 4, 8, 9], ont montré leur efficacité. La principale différence avec l'algorithme que nous proposons se situe au niveau de la représentation du modèle. La plupart des méthodes existantes traitent l'ensemble des modes équitablement. Il en résulte des ambiguïtés puisque chaque mode de déformation peut être remplacé par une combinaison linéaire des autres modes. Nous proposons au contraire de les ordonner par importance en terme d'amplitude de déformation capturée dans les images. Cette représentation permet de lever certaines ambiguïtés et introduit une estimation hiérarchique du modèle de faible rang. La sélection automatique du nombre de modes de déformation est une problématique qui est très peu étudiée. Nous proposons d'utiliser la validation croisée pour faire ce choix. En résumé, nous proposons un nouvel algorithme de reconstruction 3D basé sur le modèle de faible rang. Ce dernier utilise un modèle de caméra perspectif, sélectionne automatiquement le nombre de modes de déformation et utilise des informations a priori complémentaires favorisant la reconstruction de formes 3D plausibles. Il permet en outre l'incrustation réaliste d'objets virtuels 3D sur une vidéo d'une surface déformable.

2 Pré-requis

2.1 Notations et modèle de caméra

Les différentes instances sont exprimées en coordonnées homogènes, le symbole $\sim$ désignant l'égalité à un facteur près. La projection d'un point 3D $\mathbf{Q}_j$ par une caméra P_i est donnée par $\mathbf{q}_{i,j} \sim \mathsf{P}_i\mathbf{Q}_j$, avec $i \in [1, \cdots, n]$ les indices des images et $j \in [1, \cdots, m]$ les indices des points et où $\mathsf{P}_i = \begin{pmatrix} \overline{\mathsf{P}}_i & \mathbf{p}_i \end{pmatrix}$ représente une matrice de projection

perspective de taille (3×4). Nous définissons une carte de visibilité binaire $\mathsf{V}_{n\times m}$ dont chaque élément $v_{i,j}$ indique la présence ou non de la donnée $\mathbf{q}_{i,j}$.
L'erreur de reprojection pour un point image est la distance euclidienne $d(\mathbf{q};\widehat{\mathbf{q}})$ entre le point prédit par le modèle $\widehat{\mathbf{q}}$ et le point image correspondant $\mathbf{q}$. L'erreur de reprojection algébrique correspondante est donnée par la distance algébrique suivante :

$$d_A(\mathbf{q};\widehat{\mathbf{q}}) \stackrel{\text{def}}{=} \left\|\mathsf{S}([\mathbf{q}]_\times\, \widehat{\mathbf{q}})\right\|^2,$$

avec $[\mathbf{q}]_\times = \begin{pmatrix} 0 & -q_3 & q_2 \\ q_3 & 0 & -q_1 \\ -q_2 & q_1 & 0 \end{pmatrix}$ et $S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$.
La distance orthogonale entre un point $\mathbf{q}$ et une droite $\mathbf{l}$ est notée $d_{pl}(\mathbf{q};\mathbf{l})$ dont l'approximation algébrique est donnée par $d^2_{A\_pl} = (\mathbf{q}^\mathsf{T}\mathbf{l})^2$. Notons que les données sont normalisées afin d'améliorer les performances de l'approximation algébrique des distances.

2.2 Le modèle de faible rang

Les déformations d'un point 3D $\mathbf{S}_{i,j}$ sont modélisées en combinant une forme moyenne $\mathbf{M}_j = \begin{pmatrix} \overline{\mathbf{M}}_j & 1 \end{pmatrix}$ et l modes de déformation $b_{k,j}\mathbf{C}_{k,j}$. Ces derniers se composent d'une amplitude de déformation $b_{k,j}$ et d'une direction de déformation $\mathbf{C}_{k,j} = \begin{pmatrix} \overline{\mathbf{C}}_{k,j} & 0 \end{pmatrix}$ avec $\left\|\overline{\mathbf{C}}_{k,j}\right\| = 1$. La position d'un points 3D $\mathbf{S}^l_{i,j}$, modélisée par l modes de déformation, est alors donnée par :

$$\mathbf{S}^l_{i,j} \stackrel{\text{def}}{=} \mathsf{D}_i \left(\mathbf{M}_j + \sum_{k=1}^{l} a_{i,k} b_{k,j} \mathbf{C}_{k,j}\right), \qquad (1)$$

avec D_i la pose de la surface observée sur l'image i et $\{a_{i,k}\}_{k=1}^{l}$ sont les coefficients de forme sur l'image i. La prédiction d'un point image, c'est-à-dire la reprojection d'un point 3D $\mathbf{S}^l_{i,j}$ prédit par l'équation (1) est donnée par :

$$\mathbf{s}^l_{i,j} \sim \mathsf{P}_i \mathbf{S}^l_{i,j} = \mathsf{P}_i \mathsf{D}_i \mathbf{M}_j + \overline{\mathsf{P}}_i \overline{\mathsf{D}}_i \sum_{k=1}^{l} a_{i,k} b_{k,j} \overline{\mathbf{C}}_{k,j}, \quad (2)$$

avec $\mathsf{P}_i = \mathsf{K}_i \begin{pmatrix} \mathsf{I} & \mathbf{0} \end{pmatrix}$. Par la suite nous supposons que les caméras sont calibrées, les matrices K_i sont connues. Nous définissons également les vecteurs : $\mathbf{a}_l \stackrel{\text{def}}{=} \begin{pmatrix} a_{1,l} & \cdots & a_{n,l} \end{pmatrix}$, $\mathbf{b}_l \stackrel{\text{def}}{=} \begin{pmatrix} b_{l,1} & \cdots & b_{l,m} \end{pmatrix}$, $\overline{\mathbf{C}}^\mathsf{T}_l \stackrel{\text{def}}{=} \begin{pmatrix} \overline{\mathbf{C}}^\mathsf{T}_{l,1} & \cdots & \overline{\mathbf{C}}^\mathsf{T}_{l,m} \end{pmatrix}$ et $\overline{\mathbf{B}}^\mathsf{T}_l \stackrel{\text{def}}{=} \begin{pmatrix} \overline{\mathbf{B}}^\mathsf{T}_{l,1} & \cdots & \overline{\mathbf{B}}^\mathsf{T}_{l,m} \end{pmatrix}$.

3 Le modèle de faible rang hiérarchique

3.1 Description générale

Notre algorithme s'appuie sur la notion de "DEFORMOTION" présentée dans [10]. Les auteurs montrent que le mouvement global et les déformations d'une scène peuvent être séparés. La notion de mouvement global pour une scène déformable est étroitement liée à la notion de forme, ainsi décomposer le mouvement d'une scène en une composante rigide et des déformations est obtenue en estimant sa forme moyenne. Une fois la forme moyenne et le mouvement global estimés, les modes de déformation sont ajoutés un à un jusqu'à ce qu'un critère d'arrêt soit satisfait. L'amplitude des déformations capturées décroît au fur et à mesure que des modes sont ajoutés. Il en résulte un ensemble de modes ordonnés. Cette estimation hiérarchique permet à chaque mode de capturer l'amplitude maximale restante des données qui n'a pas été expliquée par les modes précédents. Le modèle obtenu est semblable à celui issu d'une analyse en composantes principales pour laquelle les modes sont ordonnés en fonction de leur valeur propre, c'est-à-dire en fonction de la variance des données qu'ils capturent.
Notre approche se base sur les relations suivantes, découlant du modèle de faible rang décrit par l'équation (1) :

$$\mathbf{S}^0_{i,j} = \mathsf{D}_i \mathbf{M}_j \qquad (3)$$

$$\mathbf{S}^{l+1}_{i,j} = \mathbf{S}^l_{i,j} + a_{i,l+1} b_{l+1,j} \mathbf{C}_{l+1,j}. \qquad (4)$$

En résumé, nous procédons de la manière suivante. Tout d'abord la forme moyenne ainsi que chaque déplacement D_i recalant globalement la surface au référentiel monde sont estimés par l'équation (3) ; ce qui constitue le mode 0 ou la composante rigide du modèle. Ensuite, chaque mode est triangulé[1] itérativement, c'est-à-dire que le $(l+1)^{\text{ème}}$ mode de déformation $\{b_{l+1,j}\mathbf{C}_{l+1,j}\}_{j=1}^{m}$ ainsi que les coefficients de forme associés $\{a_{i,l+1}\}_{i=1}^{n}$ sont évalués, à partir de l'équation (4). Une fonction de coût constituée de l'erreur de reprojection comme terme de données et d'informations *a priori* (décrites ci-dessous) est minimisée à chaque étape.

Informations *a priori*. En plus de l'information *a priori* implicitement utilisée sur l'ordonnancement des modes, nous utilisons explicitement deux informations *a priori* supplémentaires inspirées de [8]. La première contraint les variations temporelles de la surface. Cette hypothèse est valide si la surface observée ne se déforme pas « trop »entre deux images. Elle s'écrit :

$$\mathcal{E}_{as}(\mathbf{a}_{l+1}) \stackrel{\text{def}}{=} \left\|\Delta \mathbf{a}_{l+1}\right\|^2,$$

avec Δ un opérateur de différence finie approchant la dérivée première. La deuxième contrainte est sur la forme de la surface observée. Elle est basée sur l'observation que des points proches sur la forme moyenne le sont aussi après déformation, c'est-à-dire après l'ajout d'un mode. Cette hypothèse est valide dans le cadre de surfaces continues présentant des déformations lisses. Afin de construire cette pénalité, nous estimons la proximité entre chaque point de la forme moyenne :

$$\varphi_{j,g} \stackrel{\text{def}}{=} \rho\left(d^2\left(\mathbf{M}_j, \mathbf{M}_g\right)\right),$$

[1] Puisque le mouvement global de la surface est connu à cette étape, nous appelons « triangulation »l'estimation d'un mode.

avec ρ un noyau à support local[2]. La contrainte de surface s'écrit :

$$\mathcal{E}_{bs}(\mathbf{B}_{l+1}) \stackrel{\text{def}}{=} \sum_{j=1}^{m}\sum_{g=1}^{m} \varphi_{j,g}^2 \left\| \overline{\mathbf{B}}_{l+1,j} - \overline{\mathbf{B}}_{l+1,g} \right\|^2 = \left\| \Omega \overline{\mathbf{B}}_{l+1} \right\|^2, \quad (5)$$

où Ω est une matrice extrêmement creuse ayant $3m$ colonnes et dont le nombre de lignes est 3 fois le nombre d'éléments non nuls de $\{\varphi_{j,g}\}_{j,g=1}^{m,m}$.

Ambiguïtés. Si les modes sont estimés simultanément comme c'est le cas pour la plupart des méthodes existantes, il existe alors l^2 degrés d'ambiguïté : chaque mode peut être remplacé par une combinaison linéaire des autres modes. Dans notre approche, le mode $l+1$ est conditionné par l'estimation des l modes précédents, introduisant une seule ambiguïté par mode. En effet, les produits $\mathbf{a}_l\mathbf{b}_l^\mathsf{T}$ (présents dans l'équation (4)) entre les coefficients de forme et les amplitudes de déformation, peuvent être redéfinis tel que $\forall \nu \in \mathbb{R}^* \ \mathbf{a}_l\mathbf{b}_l^\mathsf{T} = (\nu\mathbf{a}_l)(\frac{1}{\nu}\mathbf{b}_l^\mathsf{T})$: ν constitue l'échelle du mode. Il existe également dans tous les cas une transformation Euclidienne indéterminée entre chaque D_i et la forme moyenne.

3.2 Estimation de la forme moyenne

La première étape de notre algorithme est de séparer le mouvement global des déformations de la surface. Nous avons vu en §3.1 que ceci peut être réalisé en estimant la forme moyenne c'est-à-dire la composante rigide de la surface. Pour cela, l'erreur de reprojection suivante est minimisée :

$$\min_{\{\mathbf{M}_j\}_{j=1}^m, \{\mathsf{D}_i\}_{i=1}^n} \sum_{i=1}^{n}\sum_{j=1}^{m} v_{i,j}\ d^2(\mathbf{q}_{i,j}, \mathsf{P}_i\mathsf{D}_i\mathbf{M}_j). \quad (6)$$

C'est un problème d'estimation de la structure à partir du mouvement[3] pour des caméras calibrées. Il est résolu en utilisant des techniques standards [6], comprenant notamment l'ajustement de faisceaux. Notons qu'aucune information *a priori* n'est requise ici puisque la reconstruction 3D de surfaces rigides est habituellement bien posée. Si les paramètres internes des caméras sont inconnus, elles peuvent être calibrées à partir d'une région rigide de la scène observée, par exemple l'arrière plan.

3.3 Triangulation d'un mode

Trianguler le mode $l+1$ revient à minimiser l'erreur suivante :

$$\min_{\mathbf{a}_{l+1}, \overline{\mathbf{B}}_{l+1}} \sum_{i=1}^{n}\sum_{j=1}^{m} d^2(\mathbf{q}_{i,j}, \mathbf{s}_{i,j}^{l+1}) + \lambda_a \left\|\Delta \mathbf{a}_{l+1}\right\|^2 + \lambda_b \left\| \Omega \overline{\mathbf{B}}_{l+1} \right\|^2. \quad (7)$$

C'est un problème d'optimisation non-linéaire en raison des produits entre les coefficients de forme, les amplitudes et les directions de déformation, et l'emploi de distances Euclidiennes pour comparer les données images. Nous procédons en deux étapes : tout d'abord, les informations *a priori* ne sont pas prises en compte et une estimation initiale du modèle est calculée. Celle-ci est ensuite raffinée par minimisation non-linéaire de la fonction de coût complète (7).

L'estimation initiale des paramètres $\mathbf{a}_{l+1}, \mathbf{b}_{l+1}, \mathbf{C}_{l+1}$ est obtenue par approximation de l'erreur de reprojection : nous montrons que les directions de déformation $\mathbf{C}_{l+1}$ peuvent être calculées indépendamment les unes des autres et indépendamment des autres inconnues. Ensuite, une fois les directions de déformation estimées, les coefficients de forme $\mathbf{a}_{l+1}$ ainsi que les amplitudes de déformation $\mathbf{b}_{l+1}$ sont évalués.

Initialisation des directions.

Séparation des problèmes. Les directions de déformation associées à chaque mode peuvent être estimées de manière indépendante. Pour cela, l'erreur de reprojection est réécrite en utilisant des distances point-droite. La combinaison des équations (2) et (4), conduit à :

$$\mathbf{s}_{i,j}^{l+1} \sim \mathsf{P}_i\mathbf{S}_{i,j}^l + a_{i,l+1}b_{l+1,j}\overline{\mathsf{P}}_i\overline{\mathsf{D}}_i\overline{\mathbf{C}}_{l+1,j}. \quad (8)$$

Cette équation représente un point image paramétré par sa position $a_{i,l+1}b_{l+1,j}$ le long d'une droite, passant par le point $\mathbf{s}_{i,j}^l = \mathsf{P}_i\mathbf{S}_{i,j}^l$ et avec vecteur directeur $\overline{\mathsf{P}}_i\overline{\mathsf{D}}_i\overline{\mathbf{C}}_{l+1,j}$. En remplaçant les points projetés définis par l'équation (8) dans chaque terme de l'erreur de reprojection on obtient :

$$\min_{\mathbf{a}_{l+1}, \overline{\mathbf{B}}_{l+1}} \sum_{i=1}^{n}\sum_{j=1}^{m} v_{i,j}\ d^2(\mathbf{q}_{i,j}, \mathbf{s}_{i,j}^l + a_{i,l+1}b_{l+1,j}\overline{\mathsf{P}}_i\overline{\mathsf{D}}_i\overline{\mathbf{C}}_{l+1,j}). \quad (9)$$

Afin de rendre notre problème indépendant des coefficients de forme et des amplitudes de déformation, les distances point-point sont remplacées par des distances point-droite d_{pl}^2, ce qui donne :

$$\min_{\overline{\mathbf{C}}_{l+1}} \sum_{i=1}^{n}\sum_{j=1}^{m} v_{i,j}\ d_{pl}^2(\mathbf{q}_{i,j}, \mathbf{l}_{i,j}^{l+1}) \quad \text{avec} \quad \mathbf{l}_{i,j}^{l+1} \stackrel{\text{def}}{=} \mathbf{s}_{i,j}^l \times (\mathsf{P}_i\mathsf{D}_i\mathbf{C}_{l+1,j}).$$

Dans l'équation ci-dessus, chaque direction $\overline{\mathbf{C}}_{l+1,j}$ de $\overline{\mathbf{C}}_{l+1}$ est indépendante. Évaluer les directions peut être résolue par m sous problèmes :

$$\min_{\overline{\mathbf{C}}_{l+1,j}} \sum_{i=1}^{n} v_{i,j}\ d_{pl}^2(\mathbf{q}_{i,j}, \mathbf{l}_{i,j}^{l+1}). \quad (10)$$

Estimation linéaire. La première étape permettant d'estimer les directions de déformation consiste à approximer la fonction de coût (10) pour aboutir à un problème d'optimisation linéaire au sens des moindres carrés. Pour cela les distances Euclidiennes sont remplacées par des distances algébriques $d_{A\_pl}^2 = \left(\mathbf{q}_{i,j}^\mathsf{T}\left[\mathbf{s}_{i,j}^l\right]_\times \overline{\mathsf{P}}_i\overline{\mathsf{D}}_i\overline{\mathbf{C}}_{l+1,j}\right)^2$. L'évaluation de $\overline{\mathbf{C}}_{l+1,j}$ sous la contrainte $\left\|\overline{\mathbf{C}}_{l+1,j}\right\| = 1$ est obtenue par SVD de la matrice suivante :

$$\begin{pmatrix} v_{1,j}\ \mathbf{q}_{1,j}^\mathsf{T}\left[\mathbf{s}_{1,j}^l\right]_\times \overline{\mathsf{P}}_1\overline{\mathsf{D}}_1 \\ \vdots \\ v_{n,j}\ \mathbf{q}_{n,j}^\mathsf{T}\left[\mathbf{s}_{n,j}^l\right]_\times \overline{\mathsf{P}}_n\overline{\mathsf{D}}_n \end{pmatrix}.$$

[2]Nous utilisons dans nos expériences un noyau Gaussien tronqué.

[3]"Rigid Structure-from-Motion" en anglais.

Les lignes correspondant à des données manquantes, c'est-à-dire pour lesquelles $v_{i,j} = 0$, sont enlevées de la matrice. Notons qu'un point j doit être visible dans au moins deux vues.

Raffinement non-linéaire. La deuxième étape consiste à parfaire l'estimation initiale de chaque $\overline{\mathbf{C}}_{l+1,j}$ par optimisation non-linéaire des m équations (10) en utilisant l'algorithme Levenberg-Marquardt. Cette étape est très peu coûteuse en temps de calcul puisque chaque direction possède 3 paramètres et est estimée indépendamment. Sur ces 3 paramètres, seuls 2 sont indépendants, le troisième étant fixé par la contrainte $\left\|\overline{\mathbf{C}}_{l+1,j}\right\| = 1$. Une pénalité $\left(\left\|\overline{\mathbf{C}}_{l+1,j}\right\|^2 - 1\right)^2$ est ajoutée aux fonctions de coût (10), pour chaque direction de déformation, afin d'imposer cette contrainte.

Initialisation des coefficients de forme et des amplitudes de déformation.

Principe. L'estimation des coefficients de forme $\mathbf{a}_{l+1}$ et des amplitudes de déformation $\mathbf{b}_{l+1}$ dépend de tous les paramètres puisque les points image $\mathbf{s}_{i,j}^{l+1}$ pour chaque vue i et chaque point j sont fonctions de $\mathbf{a}_{l+1}\mathbf{b}_{l+1}^{\mathsf{T}}$. Nous proposons d'exploiter l'ambiguïté 1D du modèle, décrit en §3.1, afin d'estimer linéairement les coefficients de forme et les amplitudes de déformation par normalisation successive de chaque élément contenu dans $\mathbf{a}_{l+1}$.

Les contraintes. Nous supposons que $a_{\xi,l+1} \neq 0$ pour $\xi \in [1, \cdots, n]$ et définissons $\mathbf{a}_{l+1}^{\xi} \stackrel{\text{def}}{=} \frac{\mathbf{a}_{l+1}}{a_{\xi,l+1}}$ et $\mathbf{b}_{l+1}^{\xi} \stackrel{\text{def}}{=} a_{\xi,l+1}\mathbf{b}_{l+1}$. En ne conservant que les termes de (9) dépendant de la vue ξ on obtient :

$$\min_{\mathbf{b}_{l+1}^{\xi}} \sum_{j=1}^{m} v_{\xi,j}\, d^2(\mathbf{q}_{\xi,j}, \mathbf{s}_{\xi,j}^{l} + b_{l+1,j}^{\xi}\overline{\mathsf{P}}_{\xi}\overline{\mathsf{D}}_{\xi}\overline{\mathbf{C}}_{l+1,j}). \quad (11)$$

Ce problème de minimisation peut être séparé en m sous problèmes :

$$\min_{\mathbf{b}_{l+1,j}^{\xi}} v_{\xi,j}\, d^2(\mathbf{q}_{\xi,j}, \mathbf{s}_{\xi,j}^{l} + b_{l+1,j}^{\xi}\overline{\mathsf{P}}_{\xi}\overline{\mathsf{D}}_{\xi}\overline{\mathbf{C}}_{l+1,j}), \quad (12)$$

revenant chacun à trianguler sur l'image ξ un point j appartenant à une droite. L'amplitude de déformation $b_{l+1,j}^{\xi}$ est obtenue par projection orthogonale de $\mathbf{q}_{\xi,j}$ sur la droite $\mathbf{l}_{\xi,j}^{l+1} = \mathbf{s}_{\xi,j}^{l} \times \overline{\mathsf{P}}_{\xi}\overline{\mathsf{D}}_{\xi}\overline{\mathbf{C}}_{l+1,j}$. Ceci ne peut pas toujours être résolu notamment si $v_{\xi,j} = 0$, c'est-à-dire si le point j n'est pas visible dans la vue ξ. A cette étape, nous possédons plusieurs versions de $\mathbf{b}_{l+1}$: $\left\{\mathbf{b}_{l+1}^{\xi}\right\}_{\xi=1}^{n}$ présentant chacune des données manquantes.

Estimer les $\mathbf{a}_{l+1}$ *et* $\mathbf{b}_{l+1}$. De par la relation suivante définissant $\mathbf{b}_{l+1}^{\xi}$: $\mathbf{b}_{l+1}^{\xi}a_{\eta,l+1} - \mathbf{b}_{l+1}^{\eta}a_{\xi,l+1} = 0$, le vecteur $\mathbf{b}_{l+1}$ concaténant les amplitudes de déformation peut être estimé en évaluant au préalable les coefficients contenus dans $\mathbf{a}_{l+1}$:

$$\min_{\mathbf{a}_{l+1}} \sum_{\xi=1}^{n}\sum_{\eta=1}^{n} \left\|\mathbf{b}_{l+1}^{\xi}a_{\eta,l+1} - \mathbf{b}_{l+1}^{\eta}a_{\xi,l+1}\right\|^2. \quad (13)$$

C'est un problème d'optimisation linéaire sous la contrainte $\|\mathbf{a}_{l+1}\| = 1$. Une fois le vecteur des coefficients de forme $\mathbf{a}_{l+1}$ estimé, les $\left\{\mathbf{b}_{l+1}^{\xi}\right\}_{\xi=1}^{n}$ sont denormalisés et moyennés pour aboutir à $\mathbf{b}_{l+1}$.

Raffinement non-linéaire. Cette étape consiste à minimiser l'équation (7) par l'algorithme Levenberg-Marquardt. La minimisation se fait directement sur les modes de déformation $\overline{\mathbf{B}}_{l+1}$ ce qui permet d'éviter l'emploi de contraintes pour chaque direction de déformation : $\left(\left\|\overline{\mathbf{C}}_{l+1,j}\right\| = 1\right)$. A noter que la nature extrêmement éparse des matrices mises en jeu est pris en compte afin d'accélérer le processus.

3.4 Critère d'arrêt d'ajout de modes

L'algorithme que nous venons de décrire est basé sur l'ajout itératif de modes au modèle de faible rang. Un critère définissant l'arrêt de ce processus est nécessaire. A chaque ajout d'un mode, le nombre de degré de liberté du modèle augmente et l'erreur de reprojection diminue comme l'illustrent les expérimentations. Les modèles de sélection existant, par exemple BIC ou GRIC, sont mal adaptés à notre problématique. La raison principale est qu'ils sont basés sur une distribution particulière (Gaussienne) des résidus. Dans le cadre du modèle de faible rang, les résidus doivent être interprétés différemment ; leur dépendance au bruit présent dans les images est plutôt faible. Ils proviennent en grande partie de l'écart entre le modèle de faible rang et la physique qui a engendré l'image, ce qui est difficilement modélisable paramétriquement.

Nous proposons d'utiliser la validation croisée pour sélectionner le nombre de modes. Son principe consiste à partitionner le jeu de données en un jeu d'apprentissage et un jeu de test et d'ensuite moyenner les erreurs obtenues sur les différents jeux de test. Cette approche ne suppose aucune distribution sur les résidus et reflète directement la capacité du modèle à extrapoler à de nouvelles données. Plus précisément, nous utilisons u jeux de données obtenus en désactivant certaines composantes de la matrice de visibilité. Les valeurs typiques de u sont $u \in [3, \cdots, 10]$, lors de nos expérimentations nous avons choisi $u = 4$.

Chaque jeu est un sous-ensemble des données d'entrée $\{\mathbf{q}_{i,j}\}_{i,j=1}^{n,m}$ devant garantir la reconstruction de chaque point dans chaque vue : toutes les lignes et colonnes de V contiennent au moins deux éléments non nuls. L'erreur associée à chaque jeu de test est obtenue en comparant ses données avec celles issues de la prédiction du modèle évalué sur le jeu d'apprentissage.

Le comportement typique du score de validation croisée est

de diminuer jusqu'à ce que le nombre optimal de modes soit obtenu, ensuite, le score augmente. Dans un premier temps, cette descente traduit l'incapacité du modèle à réaliser de bonnes prédictions lorsque le nombre de modes est insuffisant, il est alors trop restrictif. Une fois le nombre optimal de modes atteint, le score augmente car le modèle a alors tendance à expliquer des phénomènes indésirables (bruit sur les données, écart entre le modèle et la physique) : il devient trop flexible pour prédire de nouvelles données.

En pratique ce comportement n'est pas celui observé lorsque les informations *a priori* sont utilisées. Dans ce cas, le score de validation croisée est stable lorsque le nombre de modes est trop important : les contraintes supplémentaires inhibent les degrés de liberté superflus comme l'attestent également les résultats présentés dans [9]. Notre critère d'arrêt est donc le suivant, l'ajout de mode est stoppé lorsque le score de validation croisée augmente ou diminue faiblement par rapport à un seuil ϵ fixé à $\epsilon = 10^{-3}$ dans nos expérimentations.

4 Résultats expérimentaux

Dans cette section, nous allons présenter les différents résultats obtenus sur des données synthétiques et réelles. Les performances de reconstruction de notre algorithme sont comparées avec celles de l'algorithme TORRESANI [9]. Les performances de reconstruction de cet algorithme surpasse celles des approches de reconstruction déjà existantes basées sur le modèle de faible rang. Nous utilisons deux variantes : C2F[4] - NO PRIOR qui ne prend pas en compte les informations *a priori* décrites en §3.1 et C2F - PRIORS qui les utilisent.

4.1 Données synthétiques

Nous avons généré un jeu de données synthétiques en utilisant le modèle de visage *Candide-3* [2]. Un modèle de caméra perspectif[5] est utilisé pour projeter les formes 3D constituant la vérité terrain. La matrice de mesure est composée de $n = 70$ images et $m = 113$ points ; un bruit Gaussien de variance 2 pixels est appliqué aux points 2D. Afin de comparer les performances des différents algorithmes, l'erreur de reprojection, le score de validation croisée ainsi que l'erreur 3D sont mesurés en fonction du nombre de modes et du taux de données manquantes. Notons que l'erreur 3D est mesurée après avoir compensé une transformation de similarité entre la vérité terrain et les formes reconstruites.

La première expérience, illustrée sur la figure 1 (à gauche), traduit l'influence du nombre de modes sur l'erreur 3D. Les performances de C2F - NO PRIOR se dégradent lorsqu'un nombre de modes trop important est utilisé : au delà de 2 modes, l'erreur 3D augmente rapidement. C2F - PRIORS et TORRESANI sont beaucoup moins sensibles à ce phénomène, les erreurs 3D associées stagnent pour un nombre trop important de modes. Les déformations sont estimées plus précisément avec les approches proposées qu'avec TORRESANI lorsque le nombre de mode est optimal : l'erreur 3D est respectivement de $0,17\%$, $0,2\%$, $0,48\%$ pour C2F - NO PRIOR, C2F - PRIORS et TORRESANI. C2F - PRIORS offre globalement les meilleures performances de reconstruction en fonction du nombre de modes.

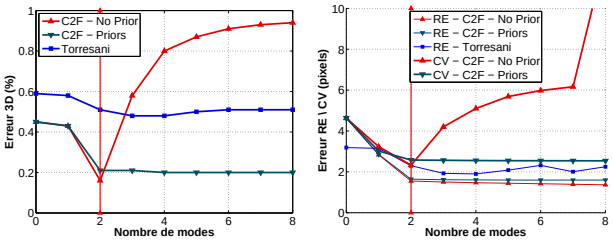

Figure 1 – *Comparaison sur des données de synthèse des algorithmes* **TORRESANI**, **C2F - NO PRIOR** *et* **C2F - PRIORS**. *A gauche : l'erreur 3D en fonction du nombre de modes. A droite : l'erreur de reprojection (RE) et le score de validation croisée (CV) en fonction du nombre de modes.*

La figure 1 (à droite) montre l'évolution de l'erreur de reprojection et du score de validation croisée en fonction du nombre de modes. Comme pressentie, l'erreur de reprojection décroît lorsque le nombre de degrés de liberté augmente tandis que le score de validation croisée se comporte sensiblement comme l'erreur 3D. Ce dernier permet de sélectionner le nombre optimal de modes pour C2F - NO PRIOR alors que pour C2F - PRIORS ce nombre est légèrement sous-estimé. Il n'y a pas cependant de dégradation significative de la forme 3D reconstruite : l'erreur 3D optimale est de $0,2\%$ et celle associée au mode selectionné est de $0,21\%$. Le taux de succès de la validation croisée sur cette expérience, pour l'algorithme C2F - NO PRIOR, c'est-à-dire sa capacité à sélectionner le nombre optimal de modes, est de 94%. Ce taux est obtenu sur 100 tirages des jeux de test et d'apprentissage. Ils prouvent la pertinence de la validation croisée pour la selection du nombre de modes. Ceci est d'autant plus vrai que le nombre de modes est sous ou sur estimé de 1 en général.

4.2 Données réelles

La vidéo du papier. Cette vidéo est composée de 203 images de taille 720×576. L'algorithme de recalage proposé dans [5] est utilisé pour suivre les déformations de la surface, 140 correspondances de points sont ainsi engendrées. Les deux algorithmes proposés : C2F - NO PRIOR et C2F - PRIORS, sélectionnent respectivement 0 et 3 mode(s), l'erreur de reprojection est respectivement de

[4]"Coarse to Fine" en anglais.

[5]Nous notons que l'algorithme TORRESANI requiert un modèle de caméra orthographique (comme la plupart des algorithmes de reconstruction 3D basés sur le modèle de faible rang). L'objectif est ici de démontrer que l'algorithme proposé est bien mieux adapté pour estimer les déformations d'une surface à partir de vidéos prises par des caméras réelles pour lesquelles l'effet perspectif est présent.

5.10 et 0.84 pixels. C2F - NO PRIOR fonctionne très mal sur cette vidéo, les déformations estimées sont biaisées : elles ne correspondent pas à la réalité. Ceci démontre que l'emploi d'informations *a priori* ne peut pas être évité. En présence de contraintes supplémentaires, la reconstruction est visuellement correcte. En outre, le score de Validation Croisée est de 1.82 pixels ce qui garantit une bonne prédiction des nouvelles données. Les résultats obtenus avec l'algorithme C2F - PRIORS sont présentés sur la figure 2. Notons que l'algorithme TORRESANI a également été testé sur cette vidéo. Les déformations ne sont que très faiblement capturées. L'erreur résiduelle est de 2,1 pixels pour 3 modes.

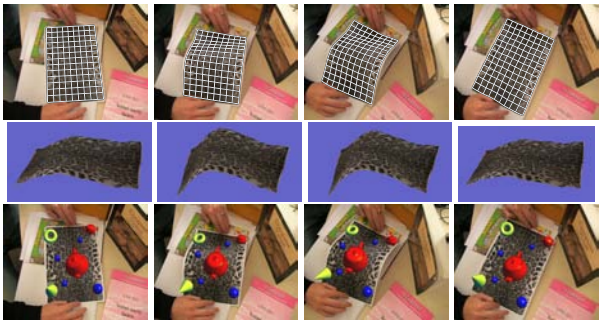

Figure 2 – *Résultats de la reconstruction sur la vidéo du papier pour l'algorithme* C2F - PRIORS. *En haut : images extraites de la vidéo ainsi que les points suivis. Au milieu : nouvelles vues synthétisées. En bas : les images augmentées.*

La vidéo « Desperate Housewives ». Nous avons extrait 100 images de taille 624×352 de la série « Desperate Housewives »représentant le personnage de Gabrielle Solis. Les déformations du visage dans les images sont suivies à l'aide d'un modèle actif d'apparence [7]. Les 68 sommets constituant l'AAM sont reconstruits avec notre algorithme, la figure 3 présente les résultats obtenus par C2F - PRIORS. C2F - NO PRIOR et C2F - PRIORS trouvent que 3 modes sont nécessaires pour capturer les déformations. Ils obtiennent respectivement 0.82 et 0.97 pixels comme erreur de reprojection et, 1.21 et 1.22 pixels pour le score de validation croisée. Ces valeurs prouvent que le modèle reconstruit est capable de prédire correctement de nouvelles données. Dans cet exemple, l'information *a priori* qu'un visage soit présent dans la vidéo est utilisée uniquement lors du suivi : notre algorithme reconstruit les déformations du visage de manière générique.

5 Conclusion

Nous proposons dans ce chapitre un algorithme de reconstruction 3D basé sur le modèle de faible rang. Ce dernier est vu comme un ensemble de modes de déformation ordonnés, estimés hiérarchiquement. Il en résulte un algorithme qui gère les données manquantes, utilise un modèle de caméra perspectif et sélectionne automatiquement le

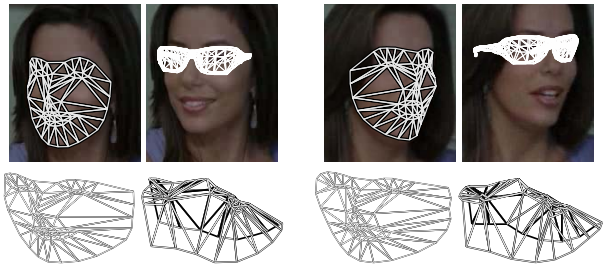

Figure 3 – *Résultats de la reconstruction sur la vidéo « Desperate Housewives ». En haut : suivi des déformations 2D par un modèle AAM et les images augmentées. En bas : reconstruction 3D des sommets du modèle.*

nombre optimal de modes de déformation par validation croisée. Nous avons également incorporé deux informations *a priori* supplémentaires dans la fonction de coût. Ces dernières améliorent grandement la qualité de la reconstruction permettant en outre l'insertion réaliste d'objets 3D virtuels sur des vidéos d'une surface déformable. Nous envisageons dans les travaux futurs de sélectionner automatiquement les poids associés aux informations *a priori*. Dans la littérature, ils sont fixés soit de manière heuristique soit par essai et erreur comme ce fut le cas lors de nos expérimentations.

Références

[1] H. Aanæs and F. Kahl. Estimation of deformable structure and motion. In *Proceedings of the Vision and Modelling of Dynamic Scenes Workshop*, 2002.

[2] J. Ahlberg. Candide 3 an Updated Parameterised Face. Technical report, Dept. of Electrical Engineering, Linköping University, Sweden, 2001.

[3] M. Brand. A direct method for 3D factorization of nonrigid motion observed in 2D. In *CVPR*, 2005.

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000.

[5] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct estimation of non-rigid registrations with image-based self-occlusion reasoning. In *ICCV*, 2007.

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. Second Edition.

[7] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004.

[8] S. Olsen and A. Bartoli. Implicit non-rigid structure-from-motion with priors. In *JMIV*, 2008.

[9] L. Torresani, A. Hertzmann, and C. Bregler. Structure-from-motion : Estimating shape and motion with hierarchical priors. *PAMI*, 2008.

[10] A. J. Yezzi and S. Soatto. Deformotion : Deforming motion, shape average and the joint registration and approximation of structures in images. *IJCV*, 2003.