Une Métrique Composite Robuste pour le Suivi de Pose du Visage utilisant un Modèle Facial Précis

Philippe Phothisane, Erwan Bigorgne, Laurent Collot, Lionel Prevost

Une Métrique Composite Robuste pour le Suivi de Pose du Visage utilisant un Modèle Facial Précis

P. Phothisane [1] E. Bigorgne [2] L. Collot [2] L. Prévost[3]

[1] ISIR - CNRS UMR 7222
Université Pierre et Marie Curie
phothisane@isir.upmc.fr

[2] EIKEO
Paris
erwan.bigorgne@majority-report.com
laurent.collot@majority-report.com

[3] LAMIA - EA 4540
Université des Antilles et de Guyane
lionel.prevost@univ-ag.fr

Résumé

Nous présentons une méthode permettant la mesure complète de la pose de la tête lors de captures vidéo monoculaires. Ce procédé utilise un modèle de visage et de texture précis. Le modèle facial, construit à partir d'une distribution de scans 3D haute résolution de visages, permet de générer des densités adaptées aux résolutions effectives. L'étape d'initialisation de la position du modèle est cruciale car elle conditionne le suivi ultérieur de la pose. Nous proposons ainsi d'utiliser une métrique composite construite sur des textures extraites de trois bases de données différentes. L'algorithme est alors capable de suivre le visage dans de vastes plages de poses avec une grande précision. Nous proposons des séquences vidéo de test dotées de vérité-terrain de pose précise et indépendante (précision de 0.1 degrés RMS).

Mots Clef

Pose, Visage, Tête, Capture de Mouvement, modèle de Texture.

Abstract

We present a method for full-motion recovery of the head pose from a monocular video input based on an accurate head model and textures. We build our face model using a distribution of high resolution 3D face scans. The cost of computations makes us select parts of this full model. To address the difficult task of initializing the model position and tracking its motion, we use a composite metric using face texture samples from three different face databases, following a positive face detection. The algorithm is subsequently able to track the head in a wide pose range with great accuracy. We also provide test video sequences with an independent accurate ground truth (with estimated RMS error of 0.1 degrees).

Keywords

Pose, Face, Head, Motion Capture, Texture Model.

1 Introduction

Le suivi de la pose visage est présent dans de nombreuses applications, de perspectives ludiques comme la réalité augmentée, à des protocoles de sécurité utilisant de la reconnaissance faciale, en passant par l'interaction homme-machine.

De nombreuses approches ont été explorées dans le suivi des mouvements de la tête. Certaines méthodes procèdent au suivi de caractéristiques faciales pour remonter à l'information de la pose. Une caractéristique faciale est la description d'une zone d'intérêt spécifique du visage. Ces zones, par exemple les coins des yeux, la pointe du nez, ou bien les commissures des lèvres sont communes à tous les visages. Elles ont en revanche une apparence et une position différentes selon les individus. Localiser et suivre ces caractéristiques sur des visages mobiles et inconnus sont de grands défis. Tong et al. ont conçu leur système de suivi [20] avec un modèle de d'apparence actif (AAM) hiérarchique [9]. Les zones caractéristiques du visage sont représentées par des ondelettes de Gabor et des profils de niveaux de gris[9]. Beaucoup d'autres systèmes de suivi sont basés sur des représentations purement bidimensionnelles des visages. Leur bon fonctionnement implique que les zones d'intérêt soient visibles, ce qui est contraignant en terme de pose du visage. Les approches tridimensionnelles ont été développées dans l'optique d'inclure des variations de rotation du visage plus larges.

Ces modèles 3D peuvent être des formes géométriques relativement simples telles que des plans[5], des cylindres[7, 22], ou bien des ellipsoïdes[8, 3]. Des modèles plus complexes utilisent des formes proches du visage humain. Certaines formes peuvent avoir une apparence relativement brute si seulement une centaine de points décrivent le visage, d'autres peuvent avoir un aspect photo-réaliste. Parmi ceux-là, citons les modèles actifs d'apparence[9, 1], ou bien les modèles déformables 3D[6, 10].

Basu et al. [3] ont développé un système visant à mesurer les déplacements de la tête, modélisée par un ellipsoïde. La

nature tridimensionnelle du modèle permet de disqualifier les points non visibles du visage pendant le processus de suivi. La texture modèle est décrite par flot optique, alors que l'estimation des paramètres optimaux de pose se fait à l'aide de l'algorithme du simplexe.

La Cascia et al. ainsi que Xiao et al. [7, 22] ont proposé une approche similaire avec un modèle de visage cylindrique. La Cascia et al. ont choisi d'optimiser les paramètres d'illumination et de pose simultanément. Xiao et al. ont appliqué une forme de l'algorithme de repondération itérative des moindres carrés [4] pour mettre à jour leur référence de texture, en considérant les probables occultations et changements d'illumination.

D'après [8], le modèle ellipsoïdal est plus adapté aux courbes du front et fournit donc de meilleurs résultats en rotation, comparativement au modèle cylindrique. Ils ont aussi développé un filtre à particule et un modèle d'apparence en ligne qui s'adapte aux changements d'apparence à courts et longs termes pendant une séquence vidéo. Dornaika et al. [10] présentent un système de suivi construit autour de Candide-3, modèle de visage déformable plus complexe, dont les points principaux coïncident avec les points remarquables du visage. Les déformations incluent des expressions faciales décrites à la main.

Notre conviction est que les modèles 3Ds extraits de véritables visages seront toujours plus adaptés au suivi de visage que des formes simplifiées. Nous proposons donc ici une méthode alternative. Bien que notre approche soit construite autour des publications [3, 7, 22], nous utilisons un modèle de visage précis extrait de scans 3D haute résolution de visages. De nos jours, un modèle aussi précis que le modèle facial de Basel (BFM) [17] n'est pas applicable dans sa forme originelle pour des implémentations de suivi de visage. Le nombre de points décrivant chaque visage implique une dimensionnalité trop élevée et des quantités de calculs trop volumineuses pour des machines communes.

Nous proposons donc de décimer le modèle initial vers un modèle allégé, tout en sauvegardant une résolution fine et des points spécifiques tels que les coins des yeux et la pointe du nez. Cette méthode diminue la haute dimensionnalité citée précédemment et son utilisation devient alors viable. Trouver la position optimale du visage de l'individu cible dans un espace de très haute dimension, sans connaissance a priori de son apparence n'est pas une tâche évidente. Nous proposons d'utiliser une métrique composite initialement introduite dans [14] pour modéliser, détecter et reconnaître les visages. Cette métrique est apprise sur les cibles inconnues, au fil de leur suivi pour converger de manière plus robuste.

Des séquences vidéo, associées à des vérités terrain permettant la validation de nos algorithmes, ont été produites par un processus de capture de mouvement. De nombreuses séquences vidéo disponibles à la communauté ont une vérité-terrain mesurée avec des capteurs magnétiques. S'ils sont faciles d'accès, ils ont aussi des défauts de précision dont la principale cause sont les inévitables interférences magnétiques qui déstabilisent et détériorent les mesures. D'autres séquences utilisent des unités de mesures inertielles qui souffrent de possibles intégrations d'erreurs. C'est pourquoi nous proposons d'utiliser une méthode précise de capture de mouvement basée sur un détecteur de formes circulaires.

Cette publication est organisée comme suit. La section II décrit toute notre méthode de suivi, en commençant part le suivi 2D, puis la construction du modèle facial décimé provenant du modèle facial dense. La méthode de Xiao et al. pour la mesure des mouvements d'un solide est décrite, ainsi que notre proposition d'amélioration, avec l'inclusion de la métrique composite. L'ajout des déformations faciales est elle aussi décrite, comme nos tests d'implémentation de l'algorithme compositionnel inverse [2]. La section III fournit des résultats de tests sur nos séquences ainsi que celles de Boston [7]. La conception et les principes de notre vérité-terrain sont expliqués, afin que la méthode puisse être facilement répliquée. Finalement, la section IV permet de conclure.

2 Méthodologie

2.1 suivi 2D

La première étape de notre système de suivi se base sur un suivi de modèle 2D rapide qui a pour objectif de gérer des variations subites d'échelle et de position. Comme une description minutieuse de notre algorithme 2D dépasse les étendues de cette publication, la présentation se fera brièvement. Dans [18], D. Ross et al. proposent une implémentation de filtre à particule qui permet un apprentissage en ligne de l'apparence de la cible : un Karhunen-Loeve Séquentiel [12] extrait incrémentalement la représentation basse dimension d'une vue affine de la cible lorsqu'elle subit des changements de pose, d'illumination et d'apparence.

Ce système se comporte bien, mais comme noté par ses concepteurs dans les conclusions de leurs études, il dérive occasionnellement de l'objet cible. En particulier, un soin particulier a été donné à l'optimisation du facteur d'oubli influant à la fois sur la stabilité et l'adaptabilité du système de suivi 2D. Nous avons donc modifié l'algorithme originel sur deux points :

– pour minimiser la dérive, le modèle d'apparence n'est plus mis à jour par des vue liées à l'état de la meilleure particule. Il n'est nourri que par des vues détectées par le détecteur de visage de Viola-Jones [21], sélectionnées par proximité spatiale et critères de taille.
– pour préserver la robustesse du système en dehors du cône de détection Viola-Jones, - dont l'angle maximal est typiquement 40°- chaque vraisemblance de particule est alors formée par un modèle mixte probabiliste qui combine sa similarité au modèle d'apparence et son autosimilarité par rapport à la dernière observation, un modèle temporaire qui évolue indépendamment pour chaque particule.

2.2 Modèle facial précis

Notre modèle facial est extrait du modèle facial de Basel (BFM) [17], qui est un modèle 3D déformable (3DMM) [6]. Ce dernier a été construit à partir de 200 scans de visages comprenant 100 hommes et 100 femmes, capturés avec des expressions neutres. L'âge moyen des individus de la base est de 25 ans, la moyenne du poids et de 67 kilogrammes, pour des sujets de 8 à 62 ans. Chaque scan facial est constitué de m = 53490 points dont la topologie est partagée. Chaque point existe donc dans chacun des visages, et est décrit par sa position 3D $(x_j, y_j, z_j)^T \in \mathbb{R}^3$, et sa couleur $(r_j, g_j, b_j)^T \in [0,1]^3, j \in \{1, ..., m\}$. Les distributions de forme et de texture sont modélisées indépendamment avec une Analyse en Composantes Principales (ACP). Le modèle facial paramétrique résultant est ainsi décrit comme :

$$M_s = (\nu_s, \sigma_s, U_s) \tag{1}$$

et

$$M_t = (\nu_t, \sigma_t, U_t) \tag{2}$$

où $\nu_{s,t}$ sont les moyennes de formes et de textures des visages (après alignement et normalisation de tous les échantillons), $\sigma_{s,t}$ sont les écarts-types associés à $U_{s,t}$ sont les bases orthonormales de formes et de textures. Cela décrit globalement le BFM, qui a été développé pour des applications de reconnaissance faciale. Les points de contrôle MPEG4 [15] sont disponibles dans une liste d'indices fournie avec la base. Ces points spécifiques auxquels nous avons rajouté d'autres points empiriquement sont explicitement non décimables dans l'algorithme de sélection suivant. Ces 108 points sont utilisés pour former une représentation faciale de type Candide 3 [10], permettant uniquement d'afficher la position courante du modèle dans le référentiel caméra. Ils n'ont pas plus de poids que les autres points du modèle.

Prendre en compte plus de 50000 points lors du suivi de la pose n'est pas nécessaire pour les résolutions sur lesquelles nous travaillons. Un tel nombre de points implique de plus des temps de calcul très longs qui ralentiraient inutilement le suivi. Il est nécessaire de sélectionner un sous-ensemble de points parmi les m points originaux. Ces m points couvrent les oreilles, le cou, et une grande partie du front. Puisque ces zones seront probablement masquées par les cheveux ou les habits dans les applications réelles, elle ont été retirées de notre modèle. Apres cette première sélection, il reste environ 26000 points (figure [1]). La décimation sur critère de densité locale de points peut alors se faire.

Le nombre de points nécessaires dépend de la résolution effective du visage suivi dans l'image courante. Par exemple, si un visage ne dépasse pas une zone englobante de 30×30 pixels, utiliser un modèle de $N = 2000$ points serait une perte de temps de calcul. Notre implémentation finale utilise une approche de régression multi-échelle (de faible à plus haute résolution) qui optimise ces temps de calculs. Les résolutions possibles de notre modèle vont de $N = 400$ à 3000 (voir figure [2]). Chaque échelle de travail utilise un modèle ayant un nombre de points adaptés à la précision requise.

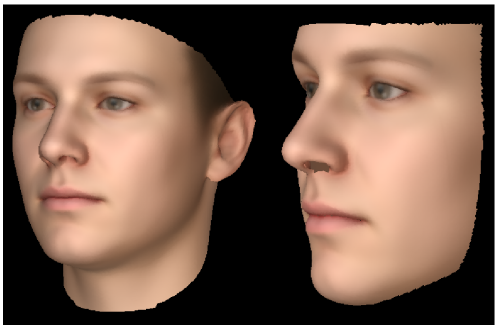

FIGURE 1 – À gauche, le modèle de Basel complet. À droite, après la première sélection.

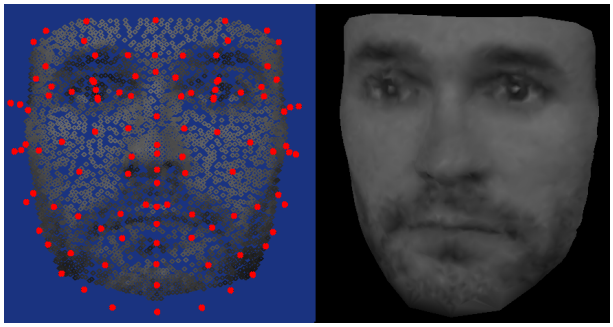

FIGURE 2 – Un exemple de projection de texture sur un modèle à 3000 points. Les points rouges sont les points de contrôle provenant de la convention MPEG4 et des points supplémentaires déterminés empiriquement, utilisés pour l'affichage du modèle pendant le suivi. Sur la droite, un rendu 3D OpenGL de la texture sur le modèle à 3000 points.

La seconde sélection de points utilise un algorithme de décimation aléatoire et itératif. À chacune des itérations, chaque point du modèle a une probabilité d'être retiré de la sélection finale. Sa probabilité de décimation dépend de sa densité locale, une densité élevée entrainant une probabilité d'être décimé élevée. À la fin, une sélection de points à densité semi-uniforme est obtenue. Passer à une plus faible résolution a pour conséquence de perdre la validité des vecteurs originels de déformation faciale. En effet, échantillonner ces vecteurs selon cette décimation détruit l'orthogonalité de la base. Pour garantir l'orthogonalité de la nouvelle base, nous avons généré aléatoirement des exemples de visages à pleine résolution en utilisant les paramètres de la distribution initiale. Pour chaque sous-résolution, les points sélectionnés sont extraits de chaque exemple généré, pour ensuite calculer une nouvelle ACP et donc une nouvelle base orthogonale.

Les distributions de textures fournies par le modèle originel n'ont pas été utilisées, car leur méthode d'acquisition et donc le rendu étant très différents de captures classiques. Pour fabriquer un modèle de texture adapté, notre modèle 3D a été placé manuellement sur de nombreux visages de différentes bases de données (FacePix [13], CMU PIE [19] et Yale face database [11]). Ainsi, notre ACP de textures est extraite de 474 textures de prises de vues classiques, en opposition à des scans de visages pris dans des conditions d'illumination uniforme très difficiles à récréer.

2.3 Méthode de suivi 3D

Optimisation au sens des moindres carrés. Reprenant la méthode de Xiao et al. dans [22], notre méthode de suivi utilise une forme des moindres carrés repondérés itérativement [4]. Nous proposons d'utiliser une autre métrique avec les notions de distance dans l'espace des caractéristiques et la distance à l'espace des caractéristiques [14].
Notre modèle 3D courant ayant N points, il est possible de mesurer le niveau de gris de chacun de ces points 3D en projetant ses coordonnées $\hat{u}_n = (x_n, y_n, z_n), n \in \{1, ..., N\}$ sur le plan image *I*. Il est supposé que cette projection ne dépend que de la longueur focale de la caméra. Pour un mouvement solide de n'importe quel point $\hat{u}_n$, sa position est mise à jour à chaque itération de l'algorithme par la matrice glisseur Q en nouvelles coordonnées $\hat{u}'_n$:

$$(\hat{u}'_n, 1) = Q.(\hat{u}_n, 1) \tag{3}$$

$$Q = \begin{pmatrix} 1 & -r_z & r_y & t_x \\ r_z & 1 & -r_x & t_y \\ -r_y & r_x & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4}$$

En combinant cette fonction transformation avec la projection on obtient la fonction *F* :

$$F(\hat{u}_n, \mu) = \begin{pmatrix} x_n - y_n r_z + z_n r_y + t_x \\ x_n r_z + y_n - z_n r_x + t_y \end{pmatrix} \cdot \frac{f_L}{-x r_y + y r_x + z + t_z} \tag{5}$$

Avec $\mu = (r_x, r_y, r_z, t_x, t_y, t_z)$, le vecteur de paramètres des mouvements solides et f_L étant la longueur focale de la caméra. $F(\hat{u}_n, \mu)$ est la projection du point courant $\hat{u}_n$ sur le plan image après une transformation solide $Q(\mu)$. Avec l'observation courante de la texture, l'objectif de l'algorithme est de trouver le vecteur de paramètres μ qui va itérativement déplacer le modèle courant vers une solution de pose de la tête. Pour cela, Xiao et al. propose de minimiser la somme de différences au carré (SSD) $E(\mu)$ entre la texture référence T et l'observation de texture courante $I(F(\hat{u}, \mu))$ avec l'algorithme de Lucas-Kanade [2].

$$E(\mu) = \sum_{n=1}^{N} (I(F(\hat{u}_n, \mu)) - T_n)^2 \tag{6}$$

Dans les cas où le visage cible est encore inconnu, pendant les premières images du suivi, tant que la pose du visage est considéré comme suffisamment de face, la référence T est une texture moyenne apprise. Les textures alors observées sont utilisées pour construire un modèle du sujet qui servira de référence après un nombre d'images S = 20 dans nos implémentations.
Selon la pose et les occultations possibles, certains points du visage ne sont pas censés être visibles. Il faut rajouter à cela le bruit dans l'image et les variations d'illumination qui peuvent perturber la texture courante et donc le suivi 3D. Pour répondre à ces problèmes, Xiao et al. utilisent une variante de la technique robuste de repondération itérative des moindres carrés (IRLS) [4]. Cette technique affecte à chaque point du modèle différents poids. Pour une pose donnée, la valeur des poids w_I dépend de la différence en niveau de gris entre la texture courante et la référence actuelle. Chacun de ces poids est d'autant plus grand que l'observation courante est proche de sa référence, et diminue plus elle s'en éloigne, en suivant une loi normale dont l'écart-type dépend de la médiane et de la distribution des niveaux de gris de la référence. Ces poids permettent de disqualifier les observations marginales. Les poids w_G dépendent pour chacun des points de la valeur absolue de son gradient, renforçant le poids des pixels de contours et des coins. Le dernier poids w_D dépend de la position du point et de sa normale définie par le modèle originel. Ce poids régularise la densité des points sur le plan image et permet de disqualifier les points invisibles en fonction de la pose du visage. Tous ces différents poids définissent alors les poids globaux w. Voir [22] pour plus de détails.
L'algorithme L.-K. cherche une solution dans le sens des moindres carrés. Nous proposons une méthode d'optimisation alternative avec la projection de la texture erreur $I(F(\hat{u}, \mu)) - T$ dans une métrique composite qui utilise la distance dans l'espace des caractéristiques (DIFS) et distance hors de l'espace des caractéristiques (DFFS) [14].

Métrique composite. Pour aisément comprendre les notions de DIFS et DFFS, imaginez une distribution gaussienne en deux dimensions. En procédant à une ACP, on obtient les deux vecteurs décrivant la distribution. L'espace des caractéristiques est toujours défini par les vecteurs correspondant aux plus grandes valeurs propres de la base. Dans notre exemple à seulement deux vecteurs, le premier vecteur décrit notre espace des caractéristiques ; alors que le second décrit son complémentaire, le non-espace des caractéristiques.
Appliqué à nos textures de visages, pour chaque résolution, nous calculons une ACP de textures provenant de bases de données de visages [13], [19], [11]. Puisque le nombre de textures disponibles est petit face à la dimensionnalité de notre espace de textures, seuls les premiers vecteurs propres contiennent de l'information pertinente. Les vecteurs suivant contenant généralement vite du bruit. Les premiers vecteurs sont sélectionnés pour construire à la fois notre DIFS et notre DFFS.
Une fois rangés, les vecteurs propres de l'ACP de texture $\{\Phi_i\}_{i=1}^{L}$ et la matrice diagonale Λ contenant les valeurs

propres associées λ_i, les K premières composantes dans $\mathbb{R}^L$ sont sélectionnées. Se forment alors la base orthogonale représentant l'espace des caractéristiques avec $\Phi_K = \{\Phi_i\}_{i=1}^{K}$ et l'espace complémentaire décrit par $\bar{\Phi}_K = \{\Phi_i\}_{i=K+1}^{L}$, qui constituent deux sous-espaces complémentaires et exclusifs. Pour n'importe quel vecteur de texture $\mathbf{x}$, nous pouvons obtenir son vecteur de paramètres $\mathbf{y} = \Phi^T(\mathbf{x} - \bar{\mathbf{x}})$. La distance de Mahalanobis s'exprime dans les termes de la somme suivante :

$$d(\mathbf{x}) = \sum_{i=1}^{K} \frac{\mathbf{y}_i^2}{\lambda_i} + \sum_{i=K+1}^{L} \frac{\mathbf{y}_i^2}{\lambda_i} \tag{7}$$

La première sommation est calculée en projetant $\mathbf{x}$ dans le principal sous-espace de dimension K. La second n'est pas calculée explicitement, les vecteurs impliqués étant trop nombreux et représentant de toute façon du bruit. La somme de ces termes est néanmoins disponible :

$$d(\mathbf{x}) = \sum_{i=1}^{L} (\mathbf{x}_i - \bar{\mathbf{x}}_i)^2 = \sum_{i=1}^{L} \mathbf{y}_i^2 \tag{8}$$

$$\epsilon^2(\mathbf{x}) = \sum_{i=K+1}^{L} \mathbf{y}_i^2 = ||(\mathbf{x} - \bar{\mathbf{x}})||^2 - \sum_{i=1}^{K} \mathbf{y}_i^2 \tag{9}$$

nous formulons un estimateur $\hat{d}(\mathbf{x})$:

$$\hat{d}(\mathbf{x}) = \sum_{i=1}^{K} \frac{\mathbf{y}_i^2}{\lambda_i} + \frac{\epsilon^2(\mathbf{x})}{\rho} \quad \text{avec} \quad \rho = \frac{1}{L-K} \sum_{K+1}^{L} \lambda_i \tag{10}$$

ρ étant la moyenne arithmétique des valeurs propres dans l'espace orthogonal, tel que l'espérance de l'estimateur soit égale à l'espérance de la distance de Mahalanobis. Au lieu d'optimiser une distance euclidienne entre notre référence T et le vecteur de texture observé $I(\hat{u})$, l'intention est de minimiser l'estimateur $\hat{d}(\mathbf{x})$:

$$\begin{aligned} \hat{d}(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}})^T \Phi_K \Lambda^{-1} \Phi_K^T (\mathbf{x} - \bar{\mathbf{x}}) + \\ (\mathbf{x} - \bar{\mathbf{x}})^T \frac{I_d - \Phi_K \Phi_K^T}{\rho} (\mathbf{x} - \bar{\mathbf{x}}) \end{aligned} \tag{11}$$

Le premier terme de (11) est la DIFS (équivalente à la distance de Mahalanobis sur les K premières composantes), tandis que le second terme est la DFFS(équivalente à la distance entre la texture courante et sa projection dans la DIFS). Au final, minimiser $\hat{d}(\mathbf{x})$, est équivalent à minimiser la nouvelle erreur $E(\mu)$, que nous utiliserons comme métrique pendant le suivi du visage.

$$E(\mu) = (I(F(\hat{u}, \mu)) - T)^T \mathbf{A} (I(F(\hat{u}, \mu)) - T) \tag{12}$$

avec

$$\mathbf{A} = \mathbf{I_d} - \Phi_K \Phi_K^T + \rho \Phi_K \Lambda^{-1} \Phi_K^T \tag{13}$$

L'algorithme L.-K. détermine l'approximation de Gauss-Newton de la matrice Hessienne

$$H = \sum_{n}^{N} \left[\nabla I_n \frac{\partial F(\hat{u}_n, 0)}{\partial \mu} \right]^T \mathbf{A} \left[\nabla I_n \frac{\partial F(\hat{u}_n, 0)}{\partial \mu} \right] \tag{14}$$

avec

$$\frac{\partial F}{\partial \mu}|_{\mu=0} \begin{pmatrix} -xy & x^2+z^2 & -yz & z & 0 & -x \\ -(y^2+z^2) & xy & xz & 0 & z & y \end{pmatrix} \cdot \frac{f_L}{z^2} \tag{15}$$

La solution fournie par Lucas-Kanade obtenue est :

$$\mu^* = -H^{-1} \sum_{n}^{N} w_n \left[\nabla I_n \frac{\partial F(\hat{u}_n, 0)}{\partial \mu} \right]^T \mathbf{A} (I(F(\hat{u}_n, 0)) - T_n) \tag{16}$$

Avec ∇I_n, la valeur du gradient de l'image pour le point n.

Modèle déformable. Pour l'instant, le modèle est un visage rigide et moyen en trois dimensions, extrait de la forme moyenne du BFM. En implémentant l'ACP de forme calculée précédemment, il est possible de déformer le modèle pour le suivi de visage. Pour des raisons évidentes, ce modèle ne peut s'adapter qu'à des visages neutres.

Les vecteurs propres de forme sont générés dans le référentiel original des visages haute-résolution. Pour une pose donnée, la matrice de rotation R qui fait tourner notre modèle de son référentiel au référentiel de la caméra conditionne les calculs permettant une adaptation à la forme. Une optimisation simultanée de la pose et de la forme implique une convergence dans un espace trop grand. Une optimisation séparée est possible. Une méthode possible est de d'abord remodeler des vecteurs de forme à $3N$ composantes $\mathbf{x}_s$, les K' premiers vecteurs propres de forme $\{\Psi_i\}, i \in \{1, ..., K'\}$ en $\tilde{\mathbf{x}}_s$ et $\left\{\tilde{\Psi}_i\right\}, i \in \{1, ..., K'\}$, matrice de taille $3 \times N$. Si le nombre choisi K' de vecteurs de forme est déterminé en fonction des valeurs propres, il faut aussi prendre en compte les temps de calculs.

Nous pouvons désormais formuler une déformation élémentaire de forme avec les vecteurs de paramètres $\alpha = (\alpha_1, ..., \alpha_{K'})$:

$$\tilde{\mathbf{x}}'_s = \tilde{\mathbf{x}}_s + R \sum_{i=1}^{K'} \tilde{\Psi}_i \alpha_i \tag{17}$$

Les coordonnées 3D et les matrices de rotations peuvent êtres séparées en lignes distinctes : $\tilde{\mathbf{x}}_s = (\tilde{x}_s, \tilde{y}_s, \tilde{z}_s)$ et $R = (R_1, R_2, R_3)$. Comme précédemment, la fonction de déformation est combinée avec fonction de projection sur le plan image :

$$G(\tilde{\mathbf{x}}_s, \alpha) = \begin{pmatrix} \tilde{x}_s + R_1 \sum_{i=1}^{K'} \tilde{\Psi} \alpha i \\ \tilde{y}_s + R_2 \sum_{i=1}^{K'} \tilde{\Psi} \alpha i \end{pmatrix} \cdot \frac{f_L}{\tilde{z}_s + R_3 \sum_{i=1}^{K'} \tilde{\Psi} \alpha i} \tag{18}$$

TABLE 1 – Erreurs de pose moyenne sur Boston.

suivi	pitch (°)	yaw (°)	roll (°)
Visage Cylindrique	4.4	5.2	2.5
Visage Ellipsoïde	3.9	4.0	2.8
M. comp. modèle précis	**3.7**	3.8	2.4
M. comp. et adapt.	4.4	**3.5**	**2.1**

Pour l'algorithme Lucas-Kanade, il faut calculer $\frac{\partial G(\tilde{\mathbf{x}},\alpha)}{\partial \alpha}$ à $\alpha = 0$:

$$\frac{\partial G}{\partial \alpha_i}|_{\alpha_i=0} = \begin{pmatrix} \tilde{z}_s R_1 \tilde{\Psi}_i - \tilde{x}_s R_3 \tilde{\Psi}_i \\ \tilde{z}_s R_2 \tilde{\Psi}_i - \tilde{y}_s R_3 \tilde{\Psi}_i \end{pmatrix} \cdot \frac{f_L}{\tilde{z}_s^2} \tag{19}$$

Pour l'adapation à la forme, la matrice Hessienne est :

$$H_m = \sum_n^N w_n \left[\nabla I_n \frac{\partial G(\tilde{\mathbf{x}}_s, 0)}{\partial \alpha} \right]^T \mathbf{A} \left[\nabla I_n \frac{\partial G(\tilde{\mathbf{x}}_s, 0)}{\partial \alpha} \right] \tag{20}$$

La solution Lucas-Kanade pour l'adaptation à la forme est :

$$\alpha^* = -H_m^{-1} \sum_n^N w_n \left[\nabla I_n \frac{\partial G(\tilde{\mathbf{x}}_s, 0)}{\partial \alpha} \right]^T \mathbf{A}(I(F(\hat{u}_n, 0)) - T_n) \tag{21}$$

Algorithme de Composition Inverse (ACI). Dans notre cas, la diminution des temps de calcul en implémentant l'ACI [2] consiste à précalculer le gradient ∇T de la référence T pour inverser les rôles des textures références et d'observation. Pour chaque nouvelle image, il faut avoir précalculé la Jacobienne $\frac{\partial F(\hat{u},0)}{\partial \mu}$ dont les termes dépendent de la position de chaque point du modèle, mais aussi le poids de régularisation w pour enfin obtenir une matrice Hessienne constante pour les prochaines itérations de l'algorithme de convergence Lucas-Kanade. Ces deux approximations impliquent que l'ACI n'est valide que dans un voisinage autour de la position initiale. Il est donc possible d'appliquer cet algorithme dans les deux schémas de régression pour les cas où la différence de pose et de position entre deux images est minime. Ces conditions spécifiques peuvent être détectées en amont par le suivi 2D.

3 Expérimentation

3.1 Base de Boston

Notre algorithme a été testé sur l'ensemble des vidéos fournies par Boston pour la pose du visage dans [7]. Le modèle du visage utilise un nombre de points entre $N = 400$ à $3,000$ dans tous les tests de cette section. Les erreurs mesurées sont calculées à partir des angles d'Euler de la matrice de rotation entre les poses courantes et les poses données par la vérité terrain. Les erreurs rapportées ne sont pas calculées selon la pose de référence de chaque séquence (désignée par Boston comme la pose de la première image), puisque dans les cas pratiques, il n'y aura que le référentiel du modèle de visage. Ces mesures contiennent donc une erreur de pose constante et différente pour chaque séquence. Les erreurs absolues sur la pose sont bornées par les résultats montrés dans la table 1.
Nous obtenons néanmoins de meilleurs résultats que les modèles cylindriques et ellipsoïdes fournis par Choi et Kim [8] (voir table 1). Notre erreur en pitch avec l'adaptation morphologique s'explique simplement. La rotation d'azimut du visage change son apparence comme certaines déformations de forme du visage de notre modèle. Exemple typique, un visage capté en contre-plongée aura visuellement un menton proéminent. Inversement, un visage pris en plongée semblera avoir un petit menton. Une convergence sur la pose peut compenser la forme et vice-versa. Cette imbrication pose/forme explique la perturbation de ces résultats. Il faudra aussi prendre en compte la précision angulaire relativement mauvaise fournie par la vérité-terrain. Il est admis que la précision est d'environ 0.5 degrés, mais étant donné le bruit donné sur les courbes de vérité-terrain, nous l'estimons à environ 3 degrés. Le bruit observé provient certainement d'interférences entre le système de mesure magnétique, le matériel électronique annexe et les structures métalliques avoisinantes [7]. Pour approfondir les expérimentations, des séquences vidéo de pose du visage associées à des vérité-terrain plus précises ont été réalisées.

3.2 Base par Capture de Mouvement

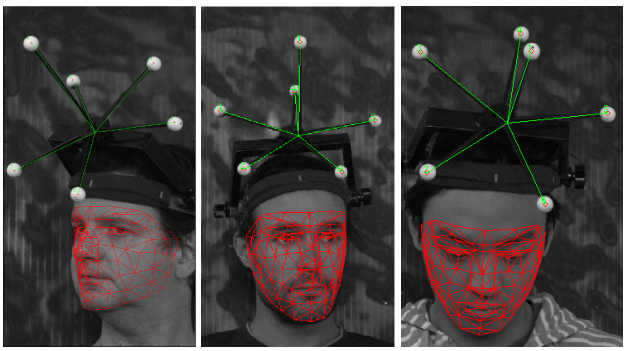

FIGURE 3 – Exemple de la capture de mouvement de notre chapeau (vert) et le suivi du visage (rouge).

Nous avons conçu un "chapeau" de capture de mouvement porté par chaque sujet du test. Le chapeau a été fabriqué pour être facilement détecté et suivi dans les séquences. La position 3D de chaque marqueur sphérique blanc a été mesurée précisément dans le référentiel du chapeau. En détectant les marqueurs dans chaque image avec une transformée de Hough circulaire, et en appariant les résultats avec les positions 3D via l'algorithme POSIT [16], (voir figure [3]), il est possible de suivre la pose du visage avec une grande précision. Les erreurs de détection des marqueurs est de l'ordre de 1 à 3 pixels, ce qui équivaut à une erreur

de 0.6 à 1.8mm. Étant donné la largeur du chapeau (environ 20cm), cela implique une erreur de 0.2 à 0.6 degrés dans le pire des cas, qui dépendent de l'attitude des sujets. L'indétermination générale est estimée à 0.1 degrés. Cette méthode évite beaucoup de problèmes de vérité-terrain de pose, comme la synchronisation des mesures et de la vidéo ou les interférences potentielles diminuant la qualité des mesures. Huit séquences vidéo comptent entre 400 et 800 images HD verticales capturées 25 fois par second. Dans chaque vidéo, les sujets ont dû d'abord fixer la caméra, puis regarder dans plusieurs directions différentes. Les séquences contiennent une grande variété de poses avec des angles de yaw et de pitch allant au-delà respectivement de 45 et 20 degrés (voir figure [4], [5]). Au total, il y a plus de 4,300 images Pour la séquence où le suivi

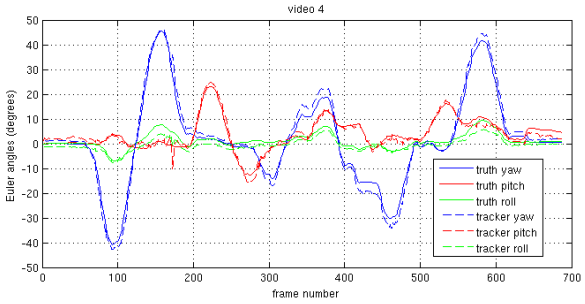

FIGURE 4 – Vérité-terrain et résultat du suivi de pose sur une vidéo test.

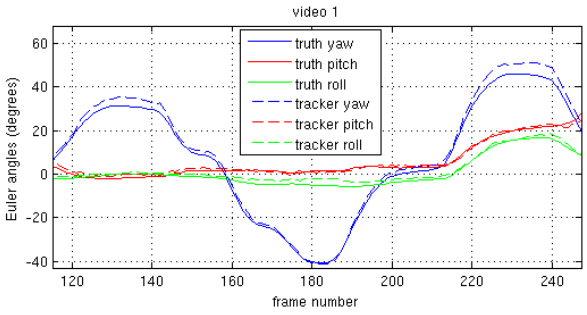

FIGURE 5 – Gros plan sur vérité-terrain et résultat du suivi de pose sur une vidéo test.

utilisant une optimisation selon les moindres carrés délivre ses meilleurs résultats, (voir table 2), l'optimisation par la métrique composite est plus précise. Comme vous pouvez le voir dans la table 3, le suivi obtient une grande précision pour des poses variées et très éloignées de la pose frontale.

4 Conclusion

Nous proposons un suivi 3D de la pose de la tête efficace utilisant un modèle de visage extrait de scans de personnes réelles en opposition à des modèles moins précis et plus géométriques. Ce modèle a été créé à l'aide des données du BFM. En décimant l'ensemble original des

TABLE 2 – Erreurs moyennes sur la vidéo 1 (Meilleurs résultats SSD). Métri . c. pour métrique composite, AM pour adaptation morphologique, et MFP pour modèle facial précis.

suivi / moyenne	pitch (°)	yaw (°)	roll (°)
SSD + AM	4.6	4.8	4.0
métri. c. MFP	3.1	3.4	3.2
métri. c. MFP + AM	**2.5**	**2.6**	**2.2**

TABLE 3 – Erreurs moyennes sur les séquences labélisées par capture de mouvement.

suivi	pitch (°)	yaw (°)	roll (°)
métri. c. MFP	4.6	3.8	2.4
métri. c. MFP + AM	**3.6**	**2.2**	**1.9**

points du modèle, un modèle plus léger a été extrait, ainsi que des déformations représentatives des visages scannés. Le processus de régression est déclenché après une détection de visage et suit de manière continue les mouvements rigides de la cible ainsi que sa morphologie à travers des séquences vidéo entières, couvrant une grande variété de poses possibles. La nature du modèle de texture permet aussi de s'adapter dans des conditions variées. Ce modèle de texture a été construit à l'aide de visages provenant de plusieurs bases de données et son implémentation dans le suivi améliore l'algorithme de Lucas-Kanade. Les bons résultats montrés sur les bases de Boston et celle de capture de mouvement montrent que notre méthode est plus précise qu'avec des modèles cylindriques ou ellipsoïdes. L'utilisation d'une méthode d'obtention de vérité-terrain de pose de la tête simple et efficace a été proposée. Cette vérité-terrain se base sur l'utilisation de capture de mouvement avec un chapeau monté de marqueurs sphériques blancs, facilement suivis par un détecteur de formes circulaires. Avec des principes triviaux, cette méthode fournit une grande précision et est facilement réplicable. Les vidéos de test incluant les vérités terrains associées sont disponibles sur demande écrite. Contactez-nous par mail pour connaître la procédure.

Références

[1] S. Baker, I. Mathews, J. Xiao, R. Gross, T. Ishikawa, and T. Kanade. Real-time non-rigid driver head tracking for driver mental state estimation. *11th World Congress Intelligent Transportation Systems*, 2004.

[2] S. Baker and I. Matthews. Lucas-kanade 20 years on : A unifying framework. *IJCV*, pages 221–255, 2004.

[3] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. *ICPR*, 3 :611–

616, 1996.

[4] M. Black. *Robust incremental optical flow*. PhD thesis, 1992.

[5] M. J. Black and Y. Yacoob. Tracking and recognizing rigid facial motions using local parametric models of image motion. *ICCV*, pages 374–381, 1995.

[6] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. *PAMI*, 2003.

[7] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination : an approach based on registration of texture-mapped 3d models. *PAMI*, 2000.

[8] S. Choi and D. Kim. Robust head tracking using 3d ellipsoidal head model in particle filter. *Pattern Recognition*, pages 241(9) :2901–2915, 2008.

[9] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *Computer Vision-ECCV*, 1998.

[10] F. Dornaika and J. Ahlberg. Fitting 3d face models for tracking and active appearance model training. *Image and Vision Computing 24*, pages 1010–1024, 2006.

[11] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. Illumination cone models for face recognition under variable lighting and pose. *PAMI*, 23 :643–660, 2001.

[12] A. Levy and M. Lindenbaum. Sequential karhunen-loeve basis extraction and its application to images. *IEEE Trans. on Image Processing*, 9 :1371–1374, 2000.

[13] G. Little, S. Krisgna, J. Black, and S. Panchanatha. A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. *Int. Conf. Acoustics, Speech and Signal Processing*, 2005.

[14] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *Early Visual Learnin*, pages 99–130, 1996.

[15] J. Ostermann. Animation of synthetic faces in mpeg-4. *Computer Animation*, pages 49–51, 1998.

[16] D. DeMenthon R. Duraiswami P. David and H. Sametl. Softposit : Simultaneous pose and correspondance determination. *IJCV*, pages 259–284, 2004.

[17] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. *6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, 2009.

[18] D.A. Ross, J. Lim, R.S. Lin, and M.H. Yang. Incremental learning for robust visual tracking. *IJCV*, 77 :125–141, 2008.

[19] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination and expression (pie) database. *FG*, 2002.

[20] Y. Tong, Q. Ji, Y. Wang, and Z. Zhu. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, 2007.

[21] P. Viola and M. Jones. Robust real-time object detector. *International Journal on Computer Vision*, pages 57 :137–154, 2004.

[22] J. Xiao, T. Kanade, and J. Cohn. Robust full-motion recovery of head by dynamic templates and re-regristeration techniques. *FG*, pages 156–162, 2002.