Suivis simultanées et robustes de visages et de gestes faciaux

Romain Hérault, Franck Davoine, Fadi Dornaika, Yves Grandvalet

Suivis simultanés et robustes de visages et de gestes faciaux

Simultaneous and robust face and facial action tracking *

HERAULT Romain[1] DAVOINE Franck[1] DORNAIKA Fadi[2] GRANDVALET Yves[1]

[1] HEUDIASYC, UMR CNRS 6599 / UTC, France
[2] Universitat Autonoma de Barcelona, Barcelone, Espagne

Université de Technologie de Compiègne
BP 20529, 60205 Compiègne cedex, France
romain.herault@hds.utc.fr

Résumé

Cet article présente une méthode permettant de suivre simultanément le visage d'une personne et les principaux traits caractéristiques de son visage, en l'occurrence les mouvements des sourcils et des lèvres. La méthode exploite un modèle d'apparence du visage composé d'un modèle géométrique 3D et un modèle de texture, ainsi qu'un modèle de dynamique. Le modèle de dynamique est estimé en ligne, à partir de la séquence d'images monoculaires. Nous considérons dans cet article deux modèles différents d'apparence, adaptés en ligne, afin d'augmenter la robustesse du suivi par rapport aux variations d'apparence du visage. Le premier repose sur une mesure robuste, permettant de pondérer les régions occultées du visage ou jugées aberrantes. Le deuxième utilise un modèle de mélange.

Mots Clef

Suivi de visages et de gestes faciaux, modèle géométrique 3D, modèle d'apparence faciale, modèle de dynamique, robustesse, mesure robuste, modèle de mélange, interaction homme-machine.

Abstract

In this work, we address a method that is able to track simultaneously head and facial actions like lip and eyebrow movements in a video sequence. In a basic framework, an adaptive appearance model is estimated online by the knowledge of the monocular video sequence. This method uses a 3D model of the face and a facial adaptive texture model. In order to increase the robustness of the tracking, we consider and compare two improved models. First, we use robust statistics in order to downweight the hidden region or outlier pixels. In a second approach, a mixture model provides better integration of occlusions. Each way will be tested separately within the basic framework.

Keywords

Head tracking, facial action tracking, 3D model, adaptive appearance model, occlusion, robust statistics, mixture model.

*Ce travail est en partie financé par le projet Behaviour, ACI Sécurité et Informatique 2004, ainsi que le réseau d'excellence PASCAL, IST-2002-506778, de la Communauté Européenne.

1 Introduction

Le suivi robuste d'un visage et de ses principaux traits caractéristiques constitue un enjeu important du domaine de la vision par ordinateur, et trouve de nombreuses applications. Citons par exemple la sécurité (surveillance, identification de personnes), ou l'interaction homme-machine. Les méthodes de suivi d'objets déformables en vision 3D reposent classiquement sur un modèle d'apparence et sur un modèle d'évolution temporelle des objets. L'apparence prise en compte peut être globale ou locale (codée par exemple sous la forme de points d'intérêt, ou de contours). Le modèle d'apparence sert à explorer l'espace de recherche considéré, à l'aide de méthodes de mise en correspondance maximisant un critère de similarité.

1.1 Travaux antérieurs

Les modèles actifs d'apparence ont été proposés comme outils puissants pour l'analyse des visages [4, 11]. Ils peuvent être utilisés pour suivre des visages dont à la fois la forme, la pose (souvent 2D) et l'apparence varient au cours du temps. Les méthodes de suivi reposant sur de tels modèles restent cependant peu robustes lorsque les conditions d'acquisition des images diffèrent de la base d'apprentissage (éclairage, propriétés de la caméra, occultation, etc.). Seuls des visages d'apparence suffisamment proche de celle des visages appartenant à la classe apprise peuvent en outre être suivis dans de bonnes conditions de précision. Dans [2], les auteurs présentent le problème général du recalage ou de l'alignement d'images, à partir d'une descente

de gradient calculée selon une approche additive ou compositionnelle. Dans [10], un modèle d'apparence basé sur un modèle de mélange est préconisé pour le suivi d'objets naturels. Il met en œuvre une estimation en ligne du modèle par un algorithme EM. Dans [6], une méthode composée de deux étapes consécutives a été développée pour le suivi de la pose 3D du visage et de ses déformations. La première étape apprend les déformations possibles des visages 3D en poursuivant des données binoculaires. La seconde étape poursuit simultanément la pose 3D et les déformations du visage en calculant le flot optique associé aux primitives suivies. Dans [3], la pose 3D du visage est estimée en recalant la texture courante du visage par rapport à une combinaison linéaire de gabarits de texture et d'illumination. Un suivi stable a été obtenu par une minimisation aux moindres carrés de l'erreur de recalage.

1.2 Présentation du travail

Dans un premier temps, nous présentons le modèle d'apparence du visage que nous utilisons pour la représentation de la structure faciale. Puis nous proposons une méthode de suivi du visage et de six gestes faciaux élémentaires, appelés aussi actions faciales, basé sur un algorithme de descente. Cette méthode étend le concept des modèles actualisés en ligne décrit dans [10, 13] au cas du suivi de la pose 3D et des mouvements des primitives du visage (mouvements des sourcils, des lèvres ...) qui sont dus, par exemple, à des expressions faciales. Elle est constituée :

– d'un modèle d'observation adaptatif, basé sur un modèle probabiliste de texture faciale apprise pendant le suivi (section 3.1),
– d'un modèle de transition adaptatif. Nous utiliserons une technique de recalage déterministe entre l'observation et la configuration courante de l'apparence (section 3.1).

Nous proposerons dans un troisième temps, deux méthodes qui améliorent la robustesse du suivi aux occultations.

– la première utilise des statistiques robustes [5] (section 3.2),
– la deuxième utilise un modèle d'observation basé sur un modèle de mélange estimé en ligne par l'algorithme EM (section 3.3).

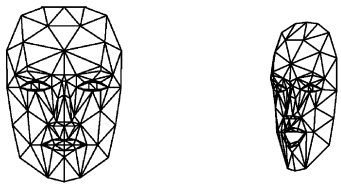
Enfin, nous présentons et comparons des résultats expérimentaux montrant notamment la robustesse aux occultations de ces deux méthodes (section 4).

2 Modèle d'apparence du visage

La modèle d'apparence du visage comprend deux composantes : le modèle de forme contenant un modèle paramétrique 3D du visage et le modèle de texture faciale. A partir de ces deux composantes, une apparence du visage peut être reconstituée.

2.1 Le modèle de forme : un modèle paramétrique 3D

Dans notre étude, nous utilisons le modèle générique 3D *Candide* [1]. Ce modèle déformable paramétrique a été initialement proposé par l'Université de Linköping pour définir des codeurs vidéo à très bas débit. En vue d'obtenir un suivi en temps réel, il remplit les conditions de facilité de manipulation et de dimension des données à stocker. Le modèle *Candide* représente la forme du visage à l'aide d'un maillage 3D composé de 200 triangles. L'identification des sommets de ces triangles $\mathbf{P}_i$ $(i = 1, \ldots, n)$ est obtenue par leurs coordonnées 3D dans un repère local. La structure 3D du modèle à un facteur d'échelle près est donnée par le vecteur $\mathbf{g}$ – la concaténation des coordonnées 3D de tous les sommets $\mathbf{P}_i$.



(a) **(b)**

FIG. 1 – **(a)** Modèle *Candide* $\bar{\mathbf{g}}$ de référence **(b)** Modèle *Candide* $\mathbf{g}$ déformé

Ce vecteur $\mathbf{g}$ de dimension $3n$ est obtenu par l'altération d'un visage de référence inexpressif :

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{S}\tau_{\mathbf{S}} + \mathbf{A}\tau_{\mathbf{A}} \quad (1)$$

où $\bar{\mathbf{g}}$ est la structure de référence du modèle. Les colonnes des matrices $\mathbf{S}$ et $\mathbf{A}$ représentent, respectivement, les unités de formes et d'actions faciales. Les vecteurs $\tau_{\mathbf{S}}$ et $\tau_{\mathbf{A}}$ codent, respectivement, quant à eux, les paramètres de déformation du visage suivant 12 modes (Tableau 1) et les paramètres de mouvement du visage suivant 6 modes (Tableau 2). L'ensemble des unités de forme $\mathbf{S}$ fournit un moyen d'adapter le modèle 3D à la physionomie du sujet. Une unité de forme applique un déplacement (codé par un vecteur) sur un ensemble réduit de points qui régissent la largeur des yeux, la hauteur du visage, la séparation des yeux, etc. L'ensemble des unités d'action $\mathbf{A}$ fournit un moyen de reproduire sur le modèle 3D les mouvements du visage. Une unité d'action applique un déplacement sur un ensemble réduit de points qui régissent la levée de la lèvre supérieure, l'abaissement de la lèvre inférieure, l'étirement des sourcils, etc.

Ainsi, le terme $\mathbf{S}\tau_{\mathbf{S}}$ tient compte de la variabilité interpersonnes tandis que le terme $\mathbf{A}\tau_{\mathbf{A}}$ tient compte de la variabilité intra-personne. $\mathbf{S}$ et $\mathbf{A}$ étant constantes dans le modèle, $\tau_{\mathbf{S}}$ et $\tau_{\mathbf{A}}$ codent les variations. Nous supposerons que ces deux variations sont découplées, c'est à dire que le vecteur $\tau_{\mathbf{A}}$ des expressions faciales sera sensé être representatif de l'ensemble de la population et donc faciliter l'apprentissage des expressions.

Pour une personne donnée, l'ensemble des unités de forme $\tau_{\mathbf{S}}$ est constant car il code la physionomie du visage. Dans le cadre de cette étude, le vecteur $\tau_{\mathbf{S}}$ est initialisé manuellement, en alignant la forme du modèle *Candide* sur la forme du visage cible présent dans la première image de la vidéo.

Nous demandons pour ce faire que le sujet se présente face à la caméra sans expressions lors de cette phase d'initialisation.

Id	Géométrie du modèle au repos
0	Hauteur de la tête
1	Position verticale des sourcils
2	Position verticale des yeux
3	Largeur des yeux
4	Hauteur des yeux
5	Séparation horizontale des yeux
6	Profondeur des joues
7	Profondeur du nez
8	Position verticale du nez
9	Position verticale du bout du nez
10	Position verticale de la bouche
11	Largeur de la bouche

TAB. 1 – Les 12 paramètres de déformation locale du visage, définissant le vecteur $\tau_{\mathbf{S}}$.

Id	Action sur le modèle au repos
0	Lever la lèvre supérieure
1	Abaisser la lèvre inférieure
2	Étirer horizontalement les lèvres
3	Abaisser les sourcils
4	Étirer verticalement les lèvres
5	Lever les sourcils

TAB. 2 – Les six actions faciales, définissant le vecteur $\tau_{\mathbf{A}}$.

La profondeur du visage peut être considérée comme très petite par rapport à la profondeur de la scène filmée, les effets de perspectives peuvent donc être négligés. Voilà pourquoi nous avons adopté une projection orthographique à l'échelle (perspective faible). La matrice de projection $\mathbf{M}$ de dimension 2×4 dépend des paramètres pose 3D du visage (rotations et translations) ainsi que de paramètres internes de la caméra (échelle). Nous projetons un sommet 3D du modèle *Candide* $\mathbf{P}_i = [X_i\ Y_i\ Z_i]^T \subset \mathbf{g}$ sur l'image en un point $\mathbf{p}_i = [u_i\ v_i]^T$ par l'équation suivante :

$$[u_i\ v_i]^T = \mathbf{M}\,[X_i\ Y_i\ Z_i\ 1]^T \tag{2}$$

Le vecteur d'état du modèle *Candide* est constitué de paramètres évoluant au cours du suivi : les informations de pose 3D (trois rotations et deux translations), les données caméras (échelle) et du vecteur de contrôle des gestes faciaux $\tau_{\mathbf{A}}$. Ceci est représenté par un vecteur $\mathbf{b}$ de dimension 12 :

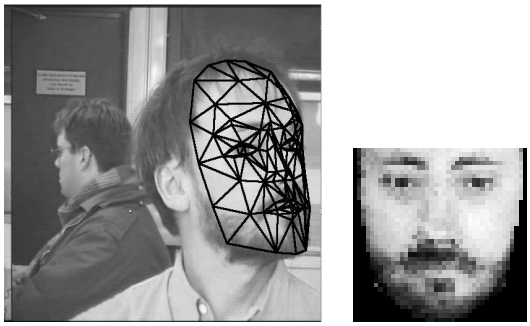
$$\mathbf{b} = [\theta_x,\ \theta_y,\ \theta_z,\ t_x,\ t_y,\ s,\ {\tau_{\mathbf{A}}}^T\ ]^T \tag{3}$$

2.2 Le modèle de texture : une texture faciale sans forme

Le modèle paramétrique 3D permet de reconstruire la forme du visage mais pas l'aspect visuel de ce dernier. Une texture faciale est plaquée sur la surface obtenue par le modèle paramétrique pour reconstituer l'apparence du visage. Un visage sera représenté sous la forme d'une texture sans forme, obtenue à partir de la texture du visage réel présent dans l'image. La projection 2D du modèle *Candide* adapté à la vue 3D du visage réel est transformée en une vue frontale 2D (modèle *Candide* 2D vu de face, de taille normalisée). La texture sans forme est obtenue à l'aide de transformations affines par morceaux appliquées à chacun des triangles du modèle projeté (cf. figure 2.b). Mathématiquement, ce processus de déformations locales appliqué à l'image originale $\mathbf{y}$ est donné par :

$$\mathbf{x}(\mathbf{b}) = \mathcal{W}(\mathbf{y}, \mathbf{b}) \tag{4}$$

où $\mathbf{x}$ représente la texture rectifiée sans forme (composée de $40\times40 = 1600$ pixels), $\mathcal{W}$ est une transformation affine par morceaux et $\mathbf{b}$ code l'état du modèle géométrique du visage. La texture rectifié est normalisée entre les valeurs -1 et 1.



(a) **(b)**

FIG. 2 – **(a)** Modèle 3D adapté à deux visages réels. **(b)** Textures rectifiées $\mathbf{x}(\mathbf{b})$ (sans forme) correspondantes.

3 Suivi de la pose 3D du visage

3.1 Méthode générale

L'objectif du suivi est l'estimation de l'état du système (dans notre cas, il est codé par le vecteur $\mathbf{b}_t$) à partir d'un ensemble d'images, $\mathbf{y}_{[t]} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$ arrivant en séquence. Le suivi possède deux composantes principales : un modèle d'observation $\mathcal{T}_t$ qui caractérise la texture faciale et un modèle de transition qui $\mathcal{D}_t$ caractérise la cinématique décrivant l'évolution de l'état entre deux observations.

Modèle d'observation adaptatif. Le modèle de texture à l'instant t, constitue le modèle d'observation $\mathcal{T}_t$. Il modélise la texture de toutes les observations jusqu'à l'instant $t-1$.

Ses paramètres varient dans le temps, c'est donc un modèle adaptatif. Au niveau de chaque image, l'observation n'est autre que la texture recalée $\mathbf{x}_\tau(\mathbf{b}_t) = \mathcal{W}(\mathbf{y}_\tau, \mathbf{b}_t)$. Le modèle de texture $\mathcal{T}_t$ est donné par une gaussienne multidimensionnelle de centre $\boldsymbol{\mu}$ et de covariance $\boldsymbol{\Sigma}^2$. On souligne que $\boldsymbol{\Sigma}^2$ est une matrice diagonale, les d composantes

de $\mathbf{b}$ étant supposées indépendantes afin de simplifier les calculs de mise à jour.
Nous supposons :

$$p(\mathbf{y}_{[t]};\mathbf{b}_t) = p(\mathbf{y}_t;\mathbf{b}_t) = p(\mathbf{x}_t) \tag{5}$$

La fonction de vraisemblance est, alors, donnée par :

$$p(\mathbf{y}_{[t]};\mathbf{b}_t) = p(\mathbf{x}_t) = \prod_{i=1}^{d} \mathbf{N}(x_i;\mu_i,\sigma_i), \tag{6}$$

où μ_i est le $i^{ème}$ élément de la diagonale de $\boldsymbol{\mu}$, σ_i est le $i^{ème}$ élément de la diagonale de $\boldsymbol{\Sigma}$, et $\mathbf{N}(x_i;\mu_i,\sigma_i)$ est une distribution gaussienne :

$$\mathbf{N}(x;\mu,\sigma) = (2\pi\sigma^2)^{-1/2}\ \exp\left[-\phi\left(\frac{x-\mu}{\sigma}\right)\right] \tag{7}$$

Avec $\phi(u) = \frac{1}{2}\,u^2$. La fonction ϕ présentée ici nous permet de trouver une densité de probabilité d'une loi normale. La fonction ϕ sera remplacée par une nouvelle fonction dans le cas d'utilisation de statistiques robustes.
Traitement des observations : Le modèle de texture n'est pas stationaire à cause, par exemple, des problèmes d'illumination. L'image courante ne suffit pas à l'estimation. Nous voulons, néanmoins, traiter toute l'information disponible afin d'estimer les paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ du modèle d'observation. Par souci de rapidité du calcul et pour économiser la mémoire, nous ne pouvons nous permettre de conserver l'ensemble des observations. Nous allons donc considérer les observation sur une fenêtre glissante. Pour ce faire, nous allons soumettre les observations à une enveloppe exponentielle. Ceci permet d' exprimer les moments de façon récursive ce qui occasionne un gain de mémoire important.
Nous supposons que $\mathcal{T}_t$ conserve toutes les observations passées sous une enveloppe exponentielle ayant un facteur d'oubli α :

$$S_t(k) = \alpha e^{\frac{k-t}{\tau}}. \tag{8}$$

avec $k < t$, $\tau = n_d/\log 2$ où n_d est la demi-vie de l'enveloppe en nombre d'images, et $\alpha = 1 - e^{-1/\tau}$. Ainsi la somme de $S_t(k)$ de $-\infty$ à 1 est égale à 1.
La log-vraisemblance de l'observation est la distance de *Mahalanobis* entre la texture recalée et l'espérance du modèle de texture $\mathcal{T}_t$:

$$\epsilon(\mathbf{b}_t) = \sum_{i=1}^{d} \phi(u_{i,t}) \tag{9}$$

où $u_{i,t} = \frac{x_i(\mathbf{b}_t)-\mu_i}{\sigma_i}$ et $\phi(u) = \frac{1}{2}\,u^2$.

Modèle de transition adaptatif. Au lieu d'utiliser une loi fixe pour le modèle de transition entre l'instant $t-1$ et l'instant t, nous profitons de l'image courante. Dans ce cas, le modèle d'évolution est donné par :

$$\mathbf{b}_t \quad = \quad \mathbf{b}_{t-1} + \Delta\mathbf{b}_t \tag{10}$$

où $\Delta\mathbf{b}_t$ code l'innovation sur les paramètres géométriques. L'estimation de (10) repose sur une technique de recalage (d'appariement) entre deux textures. L'image courante $\mathbf{y}_t$ est recalée par rapport au modèle de texture $\mathcal{T}_t$. Pour ce faire, nous minimisons l'erreur $\epsilon(\mathbf{b}_t)$.
Nous cherchons alors $\mathbf{b}_t = \mathbf{b}_{t-1} + \Delta\mathbf{b}_t$ tel que le gradient de $\epsilon(\mathbf{b}_t)$ par rapport à $\mathbf{b}$ soit nul.
Par développement limité à l'ordre 2 de la distance de *Mahalanobis*, nous obtenons :

$$\begin{aligned}\epsilon(\mathbf{b}_t+\Delta\mathbf{b}) &= \epsilon(\mathbf{b}_t) + \frac{\partial\epsilon(\mathbf{b}_t)^\intercal}{\partial\mathbf{b}}\Delta\mathbf{b} + \frac{1}{2}\Delta\mathbf{b}^\intercal + \frac{\partial^2\epsilon(\mathbf{b}_t)}{\partial\mathbf{b}\partial\mathbf{b}^\intercal}\Delta\mathbf{b} \\ &+ o(\Delta\mathbf{b}^\intercal\Delta\mathbf{b})\end{aligned} \tag{11}$$

Lorsque ϵ atteint son minimum, son gradient est nul :

$$\frac{\partial\epsilon(\mathbf{b}_t+\Delta\mathbf{b})}{\partial\mathbf{b}} = \frac{\partial\epsilon(\mathbf{b}_t)}{\partial\mathbf{b}} + \frac{\partial^2\epsilon(\mathbf{b}_t)}{\partial\mathbf{b}\partial\mathbf{b}^\intercal}\Delta\mathbf{b} + o(\Delta\mathbf{b}^\intercal\Delta\mathbf{b}) \tag{12}$$

L'approximation quadratique nous donne :

$$0 = \frac{\partial\epsilon(\mathbf{b}_t)}{\partial\mathbf{b}} + \frac{\partial^2\epsilon(\mathbf{b}_t)}{\partial\mathbf{b}\partial\mathbf{b}^\intercal}\Delta\mathbf{b} \tag{13}$$

Ce qui implique donc pour $\Delta\mathbf{b}$:

$$\Delta\mathbf{b} = -\left(\frac{\partial^2\epsilon(\mathbf{b})}{\partial\mathbf{b}\partial\mathbf{b}^\intercal}\right)^{-1}\frac{\partial\epsilon(\mathbf{b})}{\partial\mathbf{b}} \tag{14}$$

Nous recherchons les expressions de $\frac{\partial\epsilon(\mathbf{b})}{\partial\mathbf{b}}$ et $\frac{\partial^2\epsilon(\mathbf{b})}{\partial\mathbf{b}\partial\mathbf{b}^\intercal}$:

$$\epsilon(\mathbf{b}_t) \quad = \quad \sum_i \phi(u_{i,t}) \tag{15}$$

$$\frac{\partial\epsilon(\mathbf{b}_t)}{\partial\mathbf{b}} \quad = \quad \sum_i \frac{1}{\sigma_i}\phi'(u_{i,t})\frac{\partial x_i(\mathbf{b}_t)}{\partial\mathbf{b}} \tag{16}$$

$$\begin{aligned}\frac{\partial^2\epsilon(\mathbf{b}_t)}{\partial\mathbf{b}\partial\mathbf{b}^\intercal} &= \sum_i \frac{1}{\sigma_i^2}\phi''(u_{i,t})\frac{\partial x_i(\mathbf{b}_t)}{\partial\mathbf{b}}\frac{\partial x_i(\mathbf{b}_t)}{\partial\mathbf{b}}^\intercal \\ &+ \sum_i \frac{1}{\sigma_i}\phi'(u_{i,t})\frac{\partial^2 x_i(\mathbf{b}_t)}{\partial\mathbf{b}\partial\mathbf{b}^\intercal}\end{aligned} \tag{17}$$

Nous mettons en évidence l'expression des moindres carrés pondérés des équations (16) et (17) afin de retrouver une expression proche des notations de Huber [8] utilisées en statistiques robustes.

$$\begin{aligned}\frac{\partial\epsilon(\mathbf{b}_{t-1})}{\partial\mathbf{b}} &= \sum_i \frac{1}{\sigma_i}\ \ \phi'(u_{i,t-1})\ \ \frac{\partial x_i(\mathbf{b}_{t-1})}{\partial\mathbf{b}} \\ &= \sum_i \frac{1}{\sigma_i}\ \ \frac{\phi'(u_{i,t-1})}{u_{i,t-1}}\ \ u_{i,t-1}\ \ \frac{\partial x_i(\mathbf{b}_{t-1})}{\partial\mathbf{b}} \\ &= \sum_i \frac{1}{\sigma_i}\ \ u_{i,t-1}\ \ v_i\ \ \frac{\partial x_i(\mathbf{b}_{t-1})}{\partial\mathbf{b}}\end{aligned} \tag{18}$$

avec $u_{i,t-1} = \frac{x_i(\mathbf{b}_{t-1})-\mu_i}{\sigma_i}$ et $v_i = \frac{\phi'(u_{i,t-1})}{u_{i,t-1}}$.
Ce qui peut s'écrire sous forme matricielle :

$$\frac{\partial \epsilon(\mathbf{b}_t)}{\partial \mathbf{b}} = \mathbf{J}^\intercal \mathbf{\Sigma}^{-1} \mathbf{V} \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (19)$$

De même $\frac{\partial^2 \epsilon(\mathbf{b}_t)}{\partial \mathbf{b} \partial \mathbf{b}^\intercal}$ peut s'écrire :

$$\frac{\partial^2 \epsilon(\mathbf{b}_t)}{\partial \mathbf{b} \partial \mathbf{b}^\intercal} = \mathbf{J}^\intercal \mathbf{\Sigma}^{-2} \mathbf{W} \mathbf{J} + \sum_i q_i \quad (20)$$

avec $\mathbf{J}$ la matrice jacobienne $\mathbf{J}_{i,k} = \frac{\partial \mathbf{x}_i(\mathbf{b})}{\partial b_k}$ où b_k est le $k^{ème}$ élément de $\mathbf{b}$, $q_i = \frac{1}{\sigma_i} \phi'(u_{i,t-1}) \frac{\partial^2 x_i(\mathbf{b}_{t-1})}{\partial \mathbf{b} \partial \mathbf{b}^\intercal}$, et $\mathbf{\Sigma}$,$\mathbf{V}$, $\mathbf{W}$ des matrices diagonales telle que : $\mathbf{\Sigma}_{i,i} = \sigma_i$, $\mathbf{V}_{i,i} = \frac{\phi'(u_{i,t-1})}{u_{i,t-1}}$, $\mathbf{W}_{i,i} = \phi''(u_{i,t-1})$.
Nous ne savons pas estimer $\frac{\partial^2 x_i(\mathbf{b}_{t-1})}{\partial \mathbf{b} \partial \mathbf{b}^\intercal}$ donc nous ne pouvons estimer q_i. Nous faisons l'hypothèse que le dernier terme peut être négligé. En effet, $\sum_i q_i$ peut être considérée petit devant le reste de l'expression lorsque $\mathbf{b}_{t-1}$ est proche de $\mathbf{b}_t$. Cette même hypothèse est utilisée dans la méthode de Gauss-Newton.
Cette simplification permet alors d'écrire :

$$\frac{\partial \epsilon(\mathbf{b}_t)}{\partial \mathbf{b}} = \mathbf{J}^\intercal \mathbf{\Sigma}^{-1} \mathbf{V} \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (21)$$

$$\frac{\partial^2 \epsilon(\mathbf{b}_t)}{\partial \mathbf{b} \partial \mathbf{b}^\intercal} \simeq \mathbf{J}^\intercal \mathbf{\Sigma}^{-2} \mathbf{W} \mathbf{J} \quad (22)$$

L'équation (14) peut donc s'écrire :

$$\Delta \mathbf{b}_t \simeq -\left(\mathbf{J}^\intercal \mathbf{\Sigma}_t^{-2} \mathbf{W} \mathbf{J}\right)^{-1} \mathbf{J}^\intercal \mathbf{V} \mathbf{\Sigma}_t^{-2} (\mathbf{x}_t(\mathbf{b}_{t-1}) - \boldsymbol{\mu}_t) \quad (23)$$

Dans le cas de la distance de *Mahalanobis*, $\phi(u) = \frac{1}{2}u^2$. Nous avons alors :

$$\phi'(u) = u, \quad \phi''(u) = 1 \quad (24)$$

Donc,

$$\mathbf{V} = Id, \quad \mathbf{W} = Id \quad (25)$$

Ce qui simplifie l'expression (23) de $\Delta \mathbf{b}_t$:

$$\Delta \mathbf{b}_t \simeq -\left(\mathbf{J}^\intercal \mathbf{\Sigma}_t^{-2} \mathbf{J}\right)^{-1} \mathbf{J}^\intercal \mathbf{\Sigma}_t^{-2} (\mathbf{x}_t(\mathbf{b}_{t-1}) - \boldsymbol{\mu}_t) \quad (26)$$

$\Delta \mathbf{b}_t$ représente la direction de descente à suivre afin de diminuer l'erreur $\epsilon(\mathbf{b}_t)$.

Estimation de la matrice jacobienne Nous ne pouvons avoir accès à la valeur de la jacobienne car l'expression analytique de $\mathbf{x}(\mathbf{b})$ n'existe pas. Il faut donc l'estimer. L'estimation la matrice jacobienne $\mathbf{J}$ revient à estimer un gradient :

$$\mathbf{J} = \frac{\partial \mathbf{x}_t(\mathbf{b})}{\partial \mathbf{b}} \quad (27)$$

Le gradient est estimé par différentiation numérique, autour de la solution courante $\mathbf{b}_{t-1}$. La $j^{ème}$ colonne de $\mathbf{J}$ ($j = 1, \ldots, \dim(\mathbf{b})$) est calculée par différences finies :

$$\mathbf{J}_{j,t+1} \simeq \frac{\mathbf{x}_t(\mathbf{b}_t) - \mathbf{x}_t(\mathbf{b}_t + \delta\, \mathbf{q}_j)}{\delta} \quad (28)$$

où δ est un pas approprié et $\mathbf{q}_j$ est un vecteur dont les éléments sont tous nuls à l'exception du $j^{ème}$ élément qui est égal à 1. La $j^{ème}$ colonne de $\mathbf{J}$ est estimée en utilisant plusieurs pas autour de la valeur courante b_j. La valeur finale de $\mathbf{J}_{t+1}$ est obtenue en calculant une moyenne sur tous les pas utilisés :

$$\mathbf{J}_{t+1} = \sum_j \mathbf{J}_{j,t+1} \quad (29)$$

$$\mathbf{J}_{t+1} = \frac{1}{K} \sum_{k=-K/2, k\neq 0}^{K/2} \frac{\mathbf{x}_t(\mathbf{b}_t) - \mathbf{x}_t(\mathbf{b}_t + k\, \delta_j\, \mathbf{q}_j)}{k\, \delta_j} \quad (30)$$

où δ_j est le plus petit pas associé au paramètre b_j et K est le nombre de pas.
La matrice jacobienne est estimée pour chaque image. Cela présente deux avantages : (i) Un gradient variable est capable de modéliser les variations temporelles de la texture. (ii) Il est très proche du gradient réel puisqu'à chaque instant de temps il est estimé pour la configuration géométrique courante (pose 3D et actions faciales).

Algorithme de descente En pratique, la solution cherchée $\mathbf{b}_t$ est calculé par :

$$\mathbf{b}_t = \mathbf{b}_{t-1} + \rho \Delta \mathbf{b}_t \quad (31)$$

Le pas ρ est estimé de façon itérative par l'algorithme de la recherche d'or [12].

Mise à jour du modèle d'observation. Lorsque la texture $\mathbf{b}_t$ de l'image courante est disponible, la texture est actualisée et utilisée pour le suivi dans l'image suivante. Les moments sont soumis au même traitement que les observations. Nous leur appliquons une enveloppe exponentielle. Ceci permet leur expression de façon récursive :

$$\hat{M}_{j,t} = \alpha x_t^j + (1-\alpha) \hat{M}_{j,t-1} \quad (32)$$

La moyenne et la variance du modèle sont définies grâce aux moments :

$$\mu_t = M_{1,t}, \quad \sigma_t^2 = M_{2,t} - \mu_t^2 \quad (33)$$

Les vecteurs de texture $\boldsymbol{\mu}$ et $\mathbf{\Sigma}$ sont alors actualisés avec les équations suivantes :

$$\hat{\mu}_t = \alpha\, x(\mathbf{b}_t) + (1-\alpha)\, \hat{\mu}_{t-1} \quad (34)$$

$$\hat{\sigma}_t^2 = \frac{\alpha}{1-\alpha} (x(\mathbf{b}_t) - \hat{\mu}_{t-1})^2 + (1-\alpha)\, \hat{\sigma}_{t-1}^2 \quad (35)$$

Afin de prévenir tout débordement de calcul nous imposons une limite basse à la variance.
Le vecteur $\boldsymbol{\mu}_0$ est initialisé à l'aide de la texture $\mathbf{x}(\mathbf{b}_0)$ extraite de la première image de la séquence vidéo. Le vecteur $\hat{\mathbf{\Sigma}}^2$ (équation 35) n'est utilisé qu'à partir du moment où le nombre d'images excède un certain seuil (typiquement de l'ordre de 40 images). Dans nos expériences, la valeur de α, constante, est choisie dans l'intervalle $[0.01, 0.10]$.

3.2 Traitement par statistiques robustes

Modèle d'observation. Les occultations et les variations importantes de l'apparence du visage seront traitées à l'aide de mesures statistiques robustes [8]. Ces perturbations ne peuvent pas être expliquées avec le modèle de texture $\mathcal{T}_t$ seul, leur influence sur la procédure d'estimation des paramètres doit être diminuée. Nous choisissons ici d'agir sur la robustesse de (i) la fonction de vraisemblance, (ii) la méthode de descente, et (iii) la mise à jour du modèle d'apparence $\mathcal{T}_t$.
Nous utilisons pour cela une fonction ψ définie comme suit [8] :

$$\psi(u) = \begin{cases} \frac{1}{2}u^2 & \text{si} \quad |u| \leq c \\ c|u| - \frac{1}{2}c^2 & \text{si} \quad |u| > c \end{cases} \quad (36)$$

où u est la valeur centrée-réduite d'un pixel de x. La constante c contrôle la proportion des valeurs aberrantes, c.à.d. la quantité des données qui sont très éloignées de la moyenne courante de l'apparence. Dans notre étude, pour la plupart des expériences, c = 3. Si $|u| > c$ est satisfait, le pixel correspondant est déclaré aberrant.
Nous intégrons ψ dans la fonction de vraisemblance à la place de ϕ : l'influence des occultations est diminuée en appliquant une loi linéaire et non plus quadratique lorsque la valeur de l'échantillon est trop éloigné du centre de la distribution gaussienne. Lorsque l'échantillon est proche du centre de la distribution, la fonction de vraisemblance reste inchangée.

Calcul de la direction de descente. L'algorithme utilisé est le même que dans le cas général. Seul le calcul de $\Delta\mathbf{b}_t$ diffère. Dans le cadre de l'utilisation des statistiques robustes, les matrices $\mathbf{V}$ et $\mathbf{W}$ permettent alors de diminuer l'influence des pixels occultés sur le calcul de la direction de gradient. Elles ne sont plus les matrices identités et ne sont donc plus simplifiables.

$$\Delta\mathbf{b}_t \simeq -\left(\mathbf{J}^\intercal\boldsymbol{\Sigma}_t^{-2}\mathbf{W}\mathbf{J}\right)^{-1}\mathbf{J}^\intercal\mathbf{V}\boldsymbol{\Sigma}_t^{-2}(\mathbf{x}_t(\mathbf{b}_{t-1}) - \boldsymbol{\mu}_t) \quad (37)$$

Les matrices $\boldsymbol{\Sigma}$, $\mathbf{V}$ et $\mathbf{W}$ gardent les mêmes définitions : $\boldsymbol{\Sigma}_{i,i} = \sigma_i$, $\mathbf{V}_{i,i} = \frac{\psi'(u_{i,t-1})}{u_{i,t-1}}$, $\mathbf{W}_{i,i} = \psi''(u_{i,t-1})$ où $u_{i,t-1} = \frac{x_i(\mathbf{b}_{t-1}) - \mu_i}{\sigma_i}$.
Nous avons :

$$\psi(u) = \begin{cases} \frac{1}{2}u^2 & \text{si} \quad |u| \leq c \\ c|u| - \frac{1}{2}c^2 & \text{si} \quad |u| > c \end{cases} \quad (38)$$

Ainsi,

$$\psi'(u) = \begin{cases} u & \text{si} \quad |u| \leq c \\ c \times signe(u) & \text{si} \quad |u| > c \end{cases} \quad (39)$$

et

$$\psi''(u) = \begin{cases} 1 & \text{si} \quad |u| \leq c \\ 0 & \text{si} \quad |u| > c \end{cases} \quad (40)$$

Nous obtenons donc :

$$\mathbf{V}_{i,i} = \begin{cases} 1 & \text{si} \quad |u_{i,t-1}| \leq c \\ \frac{c}{|u_{i,t-1}|} & \text{si} \quad |u_{i,t-1}| > c \end{cases} \quad (41)$$

et

$$\mathbf{W}_{i,i} = \begin{cases} 1 & \text{si} \quad |u_{i,t-1}| \leq c \\ 0 & \text{si} \quad |u_{i,t-1}| > c \end{cases} \quad (42)$$

Mise à jour du modèle d'observation. Une fois la solution $\mathbf{b}_t$ disponible, la texture correspondante $\mathbf{x}$ peut être utilisée pour la mise à jour des paramètres de l'apparence. Pour les pixels qui ne sont pas déclarés aberrants, les mises à jour sont données par les équations (34) et (35) ; pour les autres, la moyenne et la variance correspondante ne sont pas mises à jour pour ne pas détériorer le modèle de texture. Une occultation est déclarée si le pourcentage de pixels détectés comme aberrants excède un certain seuil, ceci pourra être utilisé lors de la reconnaissance des gestes par un algorithme d'apprentissage.

3.3 Traitement par modèle de mélange

Les modèles de mélanges peuvent facilement intégrer des composantes différentes. Ils utilisent pour se faire un mélange de lois où chaque densité de probabilité représente une population distincte. Nous pouvons donc disposer d'un modèle qui explique à la fois la texture et les occultations. Jepson et al. [9] et Hasler et al. [7] ont proposé leur utilisation dans le cadre de l'appariement de blocs. Notre deuxième implémentation est basée sur cette démarche appliquée au suivi de la position du visage et des gestes faciaux.

Modèle d'observation. L'observation $\mathbf{x}_t$ subit un changement continu et lent lorsque le sujet modifie la position de sa tête dans l'espace ou lorsqu'il change d'expression faciale. Il n'en est pas de même avec les occultations du visage, les modifications de l'apparence (le sujet peut retirer ses lunettes), le bruit vidéo. Cela nous amène à construire un mélange de lois contenant deux composantes : une composante Stable et une composante Bruit.
La composante Stable : $\mathcal{S}$ Cette composante modélise les phénomènes stables et constants de l'observation. Sa densité de probabilité est normale $p_s(\mathbf{x}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t^2)$. Les paramètres $\boldsymbol{\mu}_t$ et $\boldsymbol{\Sigma}_t^2$ évoluent lentement dans le temps, ils constituent respectivement la moyenne et la covariance de cette loi. Cette composante représente le modèle de texture $\mathcal{T}_t$ de la méthode général sans gestion des occultations.
La composante Bruit : $\mathcal{B}$ Cette composante modélise les phénomènes instables ou soudains de l'observation tels que les occultations, le bruit. Sa densité de probabilité est uniforme : $p_b(\mathbf{x}_t)$. Comme les pixels sont normalisés dans l'intervalle $[-11]$, elle est fixée à $\frac{1}{2}$ dans notre étude.
Ces deux composantes sont combinées dans un modèle de mélange (Fig. 3). Nous obtenons pour un pixel x_t de l'observation $\mathbf{x}_t$:

$$p(x_t; m_t, \mathbf{q}_t) = m_t p_s(x_t; \mathbf{q}_t) + (1 - m_t)p_b(x_t) \quad (43)$$

où m_t représente le paramètre de proportionnalité de mélange. Les paramètres m_t et $\mathbf{q}_t$ sont estimés au cours du temps, ainsi les probabilités de mélanges indiquent l'explication des observations récentes par les différentes composantes du modèle.

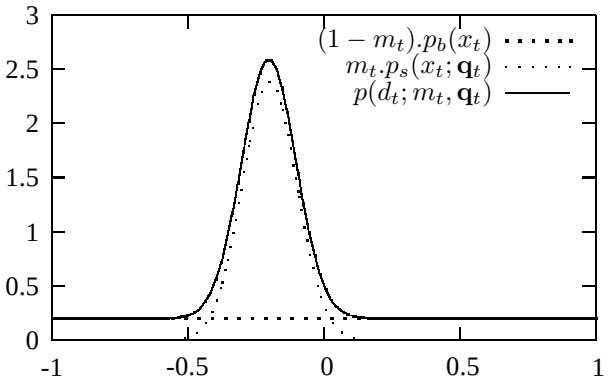


FIG. 3 – Modèle de mélange : densités de probabilité de la composante $\mathcal{S}_t$: $p_s(x_t;\mathbf{q}_t)$, de la composante $\mathcal{B}_t$: $p_b(x_t)$, mélange des lois, $p(x_t;m_t,\mathbf{q}_t)$.

Mise à jour du modèle d'observation. Nous devons pour chaque observation estimer les paramètres m_t et $\mathbf{q}_t$ de l'équation (43). Cependant, pour une estimer ces paramètres, l'algorithme EM doit pouvoir accéder à l'ensemble des observations. Jepson [10] propose une estimation en ligne des paramètres par une expression récursive de l'algorithme EM. Nous décrivons ci-après cette procédure d'estimation.

Traitement des observations : Tout comme dans la méthode générale, les observations sont pondérées par une enveloppe exponentielle. Cette pondération nous permettra de formuler les mises à jour de l'étape M de façon récursive.

$$S_t(k) = \alpha e^{\frac{k-t}{\tau}}. \tag{44}$$

Avec $k < t$, $\tau = n_d/\log 2$ où n_d est la demi-vie de l'enveloppe en nombre d'images, et $\alpha = 1 - e^{-1/\tau}$. Ainsi la somme de $S_t(k)$ de $-\infty$ à 1 est égale à 1.

L'expression du logarithme de la vraisemblance de l'observation, suivant la densité du modèle de mélange (43), est alors donné par :

$$L(\mathbf{z}_t;m_t,q_t) = \sum_{k=-\infty}^{t} S_t(k)\log p(\mathbf{z}_k;m_t,\mathbf{q}_t) \tag{45}$$

où $\mathbf{z}_t$ l'ensemble des observations : $\mathbf{z}_t = (\mathbf{x}_{-\infty},...,\mathbf{x}_{t-1},\mathbf{x}_t)$. m_t et $\mathbf{q}_t$ sont initialisés lors de la première observation, puis estimés pour les observations suivantes.

Étape E : L'étape E de l'algorithme EM consiste à calculer les probabilités d'appartenance de chaque observation x_k.

$$o_{s,t}(x_k) = \frac{m_t p_i(x_k;\mathbf{q}_t)}{p(x_k;m_t,\mathbf{q}_t)} \tag{46}$$

Comme nous avons uniquement deux composantes à notre modèle, $o_{b,t}(x_k) = 1 - o_{s,t}(x_k)$.

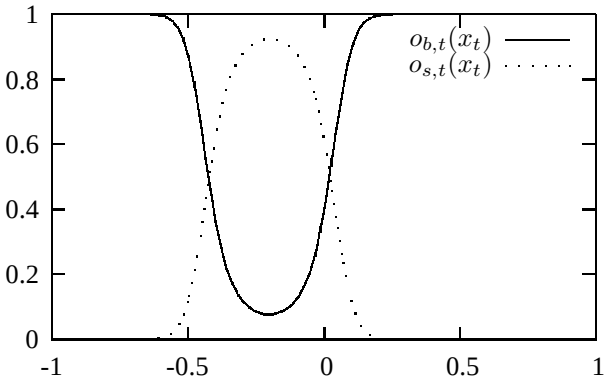


FIG. 4 – Probabilités d'appartenance : à la composante $\mathcal{S}_t$, $o_{s,t}(x_t)$, à la composante $\mathcal{B}_t$, $o_{b,t}(x_t)$.

Étape M : Une fois les probabilités d'appartenance calculées, dans l'étape M, nous pouvons mettre à jour la probabilité de mélange m_t et les moments $M_{j,t}$ d'ordre j des données pondérés par les appartenances :

$$m_t = \sum_{k=-\infty}^{t} S_t(k) o_{s,t}(x_k) \tag{47}$$

$$M_{j,t} = \sum_{k=-\infty}^{t} S_t(k) x_k^j o_{s,t}(x_k) \tag{48}$$

Il est noté qu'ici $M_{0,t} = m_{s,t}$. Ces moments permettent de calculer la moyenne et la variance de la première composante :

$$\mu_t = \frac{M_{1,t}}{M_{0,t}}, \quad \sigma_t^2 = \frac{M_{2,t}}{M_{0,t}} - \mu_t^2 \tag{49}$$

L'algorithme EM consiste à réitérer l'étape E et l'étape M à chaque nouvelle observation.

Expression récursive de l'étape M : Le calcul de la probabilité de mélange et des moments fait intervenir la probabilité d'appartenance $o_{s,t}(x_k)$. Le calcul necessite de conserver l'ensemble des observations. Nous devons donc adopter une approximation des mise à jour effectuées à l'étape M.

Dans un premier temps, nous allons exploiter l'expression de l'enveloppe $S_t(k)$:

$$S_t(k) = e^{-1/\tau} S_{t-1}(k) = (1-\alpha) S_{t-1}(k) \tag{50}$$

L'équation (48) peut alors s'écrire sous la forme :

$$M_{j,t} = S_t(t) x_t^j o_{s,t}(x_t) + \sum_{k=-\infty}^{t-1} S_t(k) x_k^j o_{s,t}(x_k)$$

$$M_{j,t} = \alpha x_t^j o_{s,t}(x_t) + (1-\alpha) \sum_{k=-\infty}^{t-1} S_{t-1}(k) x_k^j o_{s,t}(x_k) \tag{51}$$

Nous évitons le calcul de l'appartenance courante des données précédentes en l'approchant par l'appartenance obtenue lors de leur observation. Pour ce faire, nous remplaçons $o_{s,t}(x_k)$ par $o_{s,k}(x_k)$, nous obtenons ainsi une expression récursive de la mise à jour des moments :

$$\hat{M}_{j,t} = \alpha x_t^j o_{s,t}(x_t) + (1-\alpha) \sum_{k=-\infty}^{t-1} S_{t-1}(k) x_k^j o_{s,t}(x_k)$$

$$\hat{M}_{j,t} = \alpha x_t^j o_{s,t}(x_t) + (1-\alpha)\hat{M}_{j,t-1} \quad (52)$$

De la même façon, de l'équation (47) découle un approximation des probabilités de mélange par une expression recursive :

$$\hat{m}_t = \alpha o_{s,t}(x_t) + (1-\alpha)\hat{m}_{t-1} \quad (53)$$

Modèle de transition. Les occultations ne peuvent êtres expliquées par la composante $\mathcal{S}_t$ stable de notre modèle de mélange, en revanche elles sont intégrées la composante $\mathcal{B}_t$ du modèle. Les occultations et les variations importantes de l'apparence du visage ont donc pour conséquence une augmentation de la probabilité d'appartenance $o_{b,t}(x_k)$ au détriment de la probabilité d'appartenance $o_{s,t}(x_k)$. L'algorithme EM tient compte de ces probabilités pour le calcul de la fonction de vraisemblance et la mise à jour du modèle d'apparence (équation 49). Il nous reste donc à prendre en compte la gestion des occultations lors du calcul de la descente de gradient. De la même manière que pour la méthode précédente nous modifions la matrice diagonale $\mathbf{V}$, afin de diminuer l'influence des pixels aberrants sur le processus d'estimation. Cette matrice est donnée à l'instant t par :

$$\mathbf{V}_{i,i} = o_{s,t}(x_i) \quad (54)$$

Le décalage utilisé par la méthode du recalage itératif devient ainsi :

$$\Delta \mathbf{b}_t \simeq -\left(\mathbf{J}^\intercal \mathbf{\Sigma}_t^{-2} \mathbf{V} \mathbf{J}\right)^{-1} \mathbf{J}^\intercal \mathbf{V} \mathbf{\Sigma}_t^{-2} (\mathbf{x}_t(\mathbf{b}_{t-1}) - \boldsymbol{\mu}_t) \quad (55)$$

4 Résultats expérimentaux

4.1 Suivi du visage

Les figures 5 et 9 illustrent l'estimation de la pose 3D et des actions faciales associées à une séquence de 591 images (ici les images 93, 204 et 441 sont affichées). Le suivi est robuste vis-à-vis d'importants mouvements du visage et de ses actions faciales. La résolution spatiale est de 512×512 pixels. Les courbes de la figure 9 montrent l'évolution temporelle de la valeur estimée de la pose 3D ainsi que les six paramètres associés aux lèvres et aux sourcils.

4.2 Comparaison des trois méthodes

Nous avons provoqué des occultations avec des objets de deux catégories : une main dont la texture est proche de celle du visage ; une barre plastique dont la texture est éloignée de celle du visage.

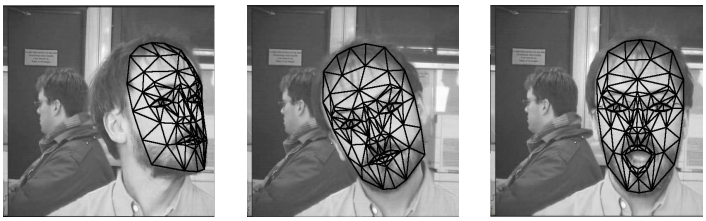
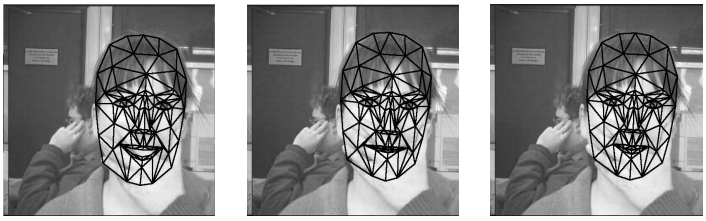


FIG. 5 – Images extraites de la séquence analysées

Occultation par un objet de texture similaire. Lorsqu'une main (texture chair) cache le visage, le suivi n'est pas affecté (Figure 6). Si la main est restée longtemps positionnée devant le visage, le modèle 3D a tendance à se déformer pour suivre la main, lorsque cette dernière se déplace, tout en restant centré sur le visage (a). Il revient en position lorsque la main quitte complètement la surface du visage. Cette déformation est moins importante avec les statistiques robustes (b) et n'apparaît pratiquement pas avec le modèle de mélange (c). La figure (10) illustre cette déformation pour le paramètre "Abaissement de la lèvre inférieure" pour les différentes méthodes. Les barres centrales délimitent le début et la fin de l'occultation.
Si la main reste devant le visage, sa texture est petit à petit intégrée au modèle de texture. Lorsque la main bouge, le modèle de dynamique peut alors interpréter ce mouvement comme un mouvement du visage, ce qui conduit à un déplacement du masque 3D. La non mise jour ou la pondération des pixels aberrants par les méthodes robustes permettent de diminuer ce phénomène.

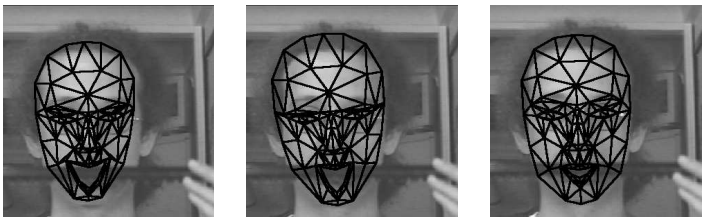


(a) Méthode générale (b) Statistiques robustes (c) Modèle de mélange

FIG. 6 – Occultation par un objet de texture similaire

Occultation par un objet de texture dissimilaire. Lorsque la barre occulte le visage, le suivi est perturbé (figure 7) : il y a une dérive de la pose de la tête et les gestes faciaux sont anormaux par la méthode générale (a), la dérive de la pose est moins importante avec les statistiques robustes (b) et seulement un petit artefact sur les gestes faciaux est visible avec le modèle de mélange (c).
La barre est correctement détectée comme occlusion par les deux méthodes robustes (figure 8) : les pixels appartenant à la barre de la texture sans forme (a) sont bien classifiés comme aberrants avec la première méthode (b) ou rejetés dans la composante Bruit avec la deuxième méthode (c). Les artefacts sur les paramètres sont aussi diminués comme le montre la figure 11a. Néanmoins avec la méthode géné-

rale et les statistiques robustes la dérive reste trop importante sur certain paramètre et conduit à une perte de la cible (figure 11b).



(a) Méthode générale (b) Statistiques robustes (c) Modèle de mélange

FIG. 7 – Occultation avec un objet de texture dissimilaire

(a) Texture faciale (b) Pixels aberrants (c) Appartenances

FIG. 8 – Texture faciale en présence de l'occultation

5 Conclusion

Dans cette étude, nous avons traité du problème de suivi du visage et des gestes faciaux dans une séquence vidéo. Nous avons présenté deux approches basées sur un modèle d'apparence globale du visage (la texture faciale) plutôt que des primitives locales. Dans la première approche, nous utilisons des statistiques robustes pour gérer les pixels aberrants dus aux occultations. Dans la deuxième, le modèle d'apparence qui était donné par une loi gaussienne est remplacé par un modèle de mélange. Ces méthodes possèdent toutes deux l'avantage d'être flexibles puisque le modèle d'apparence est appris durant la phase de suivi et non pas auparavant. Elles ne sont donc pas spécifiques à un sujet donné. Nous avons pu traiter de longues séquences vidéo pré-enregistrées (10 minutes soit 15000 images) mais aussi des acquisitions directes sur de multiples visages, dans la limite de 4 à 5 images par seconde . Un suivi de qualité, permettant d'extraire la position de la tête et les expressions faciales, a été obtenu par les deux méthodes développées, ceci même en présence d'occultations et de variations de luminosité, ce qui démontre leur efficacité.

Plusieurs améliorations peuvent cependant être envisagées. Lorsqu'il y a occultation la partie occultée est volontairement pondérée pour ne pas influer le calcul de l'erreur et de la direction de gradient. Un effet indésirable de cette technique est de rendre possible des mouvements aberrants du modèle de forme dans les zones occultées, car l'erreur n'y est plus prise en compte. Une contrainte de continuité sur les paramètres permettrait de diminuer cet effet. D'autre part, l'indépendance des pixels a été volontairement introduit pour des raisons calculatoires or les pixels occultés sont pour la plupart voisin, il serait donc intéressant de traiter les pixels par zone et non plus indépendamment.

Références

[1] J. Ahlberg. CANDIDE-3 - an updated parametrized face. Technical Report LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden, 2001.

[2] S. Baker and I. Matthews. Lucas-Kanade 20 years on : A unifying framework. *International Journal of Computer Vision*, 56(3) :221–255, 2004.

[3] M. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination : An approach based on registration of texture-mapped 3D models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4) :322–336, 2000.

[4] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6) :681–684, 2001.

[5] F. Davoine and F. Dornaika. Head and facial animation tracking using appearance-adaptive models and particle filters. In V. Pavlovic B. Kisacanin and T. S. Huang, editors, *Real-Time Vision for Human-Computer Interaction*, pages 121–140. Springer Verlag, 2005.

[6] S. Gokturk, J.-Y. Bouguet, and R. Grzeszczuk. A data-driven model for monocular face tracking. In *Proc. IEEE International Conference on Computer Vision*, Vancouver, Canada, 2001.

[7] D. Hasler, L. Sviaz, S. Susstrunk, and M. Vetterli. Outlier modeling in image matching. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 25, pages 301–315, 2003.

[8] P.J. Huber. *Robust Statistics*. Wiley, 2003.

[9] A. Jepson and M. Black. Mixture models for optical flow computation. In *CVPR93*, pages 760–761, New York City, NY, USA, 1993.

[10] A. Jepson, D. Fleet, and T.F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10) :1296–1311, 2003.

[11] I. Matthews and S. Baker. Active appearance models revisited. Technical Report CMU-RI-TR-03-02, The Robotics Institute, Carnegie Mellon University, 2002.

[12] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. Golden section search in one dimension. In *Numerical Recipes in C : The Art of Scientific Computing*, chapter 10.1. Cambridge University Press, New York, NY, USA, 1992.

[13] S. Zhou, R. Chellappa, and B. Mogghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13 :1491–1506, Nov. 2004.

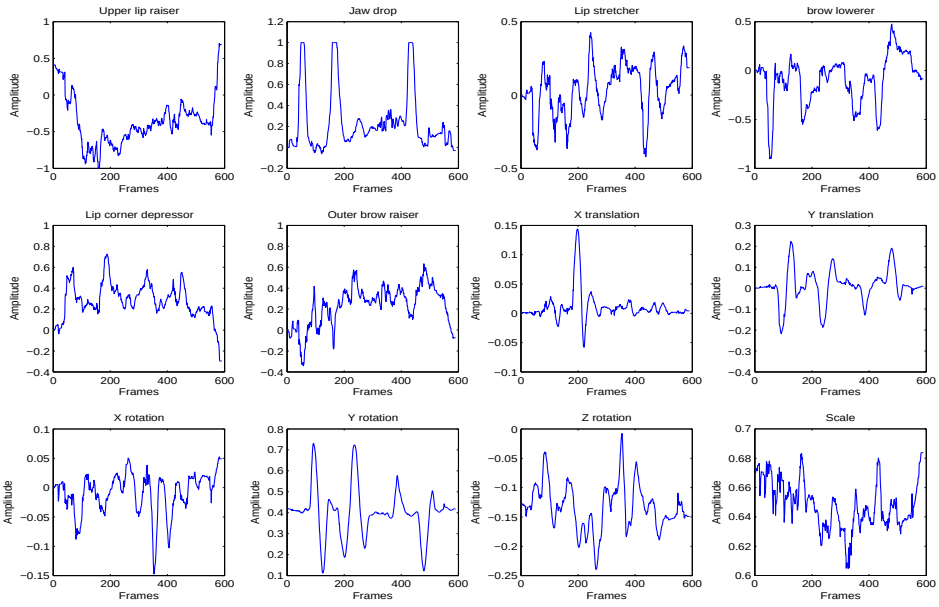


FIG. 9 – Évolution du suivi des 6 gestes faciaux et des 6 paramètres de la pose 3D au cours du temps

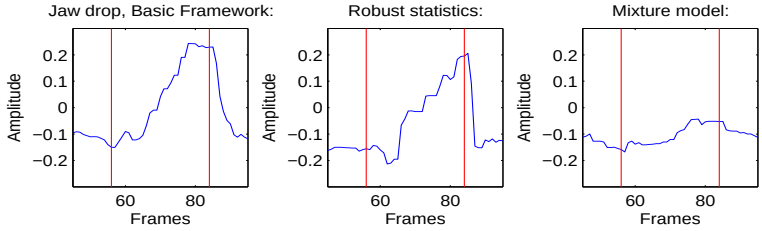


FIG. 10 – Artefacts provoqués par l'occultation du visage par un objet de texture chair

(a)

(b)

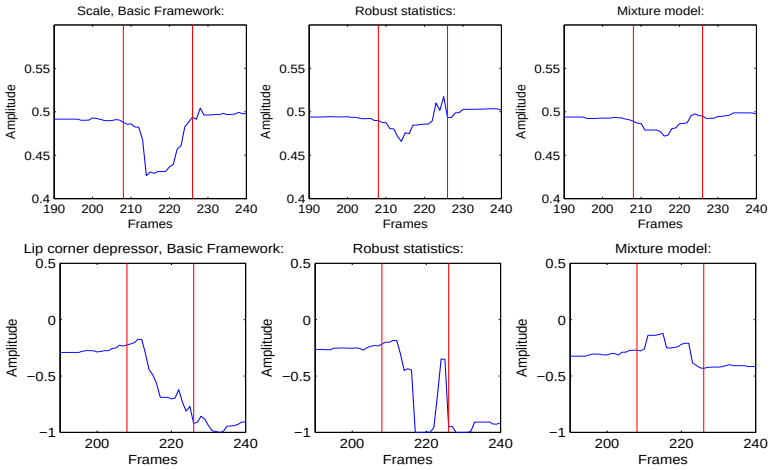


FIG. 11 – Artefacts provoqués par l'occultation du visage par un objet noir