Suivi des mouvements faciaux et de la pose 2D d'un visage

Soumya Hamlaoui, Franck Davoine,

Laboratoire HEUDIASYC, UMR - CNRS 6599 / Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex, France
Soumya.Hamlaoui@hds.utc.fr, Franck.Davoine@hds.utc.fr

Résumé – Nous considérons dans cet article le problème du suivi de la pose et des mouvements faciaux d'un visage faisant face à une caméra. Pour cela, nous proposons une approche stochastique reposant sur le filtrage particulaire où la distribution des observations est dérivée soit d'un modèle d'apparence actif, soit d'un modèle d'apparence calculé en ligne. L'évolution du système d'état est décrite par une dynamique guidées par une recherche déterministe. Le nombre de particules est ajusté aux besoins effectifs du suivi à chaque pas temporel; cet ajustement permet l'optimisation du temps de calcul du filtre. La prise en compte d'une mesure robuste permet d'augmenter la robustesse du suivi lorsque le visage est partiellement occulté. L'article se termine par la présentation de résultats expérimentaux validant l'intérêt des méthodes de suivi reposant sur les deux modèles d'observation proposés.

Abstract – In this paper, we consider the problem of tracking the global motion of a face as well as the local motion of its inner features. In this view, we propose a stochastic approach based on a particule filtering scheme. The observations distribution is derived from an active appearance model, or from an "on-line" estimated appearance model. The dynamics describing the state system evolution are guided by a deterministic research. The particles number is adjusted to the effective needs of the tracking at each time step; this adjustment allows an optimization of the computing time. We also use a robust distance measure which increases the tracking robustness when the face is partially occulted. Experimental results are presented to validate the tracking methods based on the two suggested observation models.

1 Introduction

Cet article traite de la problématique du suivi de la pose 2D ainsi que des mouvements faciaux d'un visage faisant face à une caméra, dans une séquence vidéo. La pose 2D est définie par : la position, le facteur d'échelle et l'angle de rotation du visage dans le support image. Les mouvements faciaux correspondent aux variations d'apparence (texture et forme) du visage. Nous proposons pour cela une approche basée sur l'algorithme de Condensation [4]. La distribution des observations considérée est dérivée soit d'un modèle d'apparence actif [1], soit d'un modèle d'apparence calculé en ligne. La dynamique des particules est adaptative dans le sens où elle est guidée par une recherche déterministe autour de l'hypothèse optimale prédite. Cette recherche correspond à une adaptation automatique du modèle d'apparence actif par une méthode itérative de type *descente de gradient*. Contrairement à l'algorithme de Condensation classique, à chaque pas temporel, le nombre de particules considéré est adapté aux besoins effectifs du suivi proportionnellement à une distance estimée entre l'hypothèse d'état optimale à l'instant précédent et sa prédiction à l'instant courant. Cet ajustement permet de réduire les coûts de calcul. Dans le but d'assurer la robustesse du suivi au cours d'une occultation partielle du visage, le modèle d'observation considère une mesure robuste [3]. Celle-ci permet de minimiser la contribution des pixels aberrants n'appartenant pas au visage, dans l'estimation de la vraisemblance et par conséquent dans l'approximation de la densité de probabilité a posteriori. La première section est dédiée à introduire le modèle d'apparence ainsi que l'algorithme de Condensation. Une description plus explicite de l'approche proposée est présentée dans la deuxième section, où nous détaillons le modèle dynamique ainsi que les deux modèles d'observations proposés. Nous présentons en troisième section les résultats expérimentaux obtenus.

2 Outils

2.1 Modèle d'apparence actif

Le modèle d'apparence actif [1] est une représentation statistique linéaire des variations d'apparence de la classe des visages. Cette représentation est obtenue par une Analyse en Composantes Principales (ACP) d'un ensemble de formes $\mathbf{s}$ et de textures $\mathbf{g}$:

$$\mathbf{s}^i = \mathbf{s}_m + \phi_s \mathbf{b}_s^i \qquad \mathbf{g}^i = \mathbf{g}_m + \phi_g \mathbf{b}_g^i \tag{1}$$

$\mathbf{s}_m$, $\mathbf{g}_m$ représentent la forme et la texture moyennes, ϕ_s, ϕ_g sont les vecteurs propres des matrices de covariance de forme et texture, et $\mathbf{s}^i$, $\mathbf{g}^i$ représentent respectivement une forme et une texture reconstruites. Une troisième ACP est alors appliquée sur le vecteur $\mathbf{b}^i$, obtenu par concaténation des composantes principales de forme et de texture $(\mathbf{b}_s^i, \mathbf{b}_g^i)$ correctement pondérées, afin d'obtenir le vecteur $\mathbf{c}$ du modèle d'apparence combiné : $\mathbf{b}^i = \phi_c \mathbf{c}^i$.
De nouvelles instances de forme $\mathbf{s}_{modele}$ et de texture $\mathbf{g}_{modele}$ peuvent alors être générées à partir d'un vecteur $\mathbf{c}$:

$$\mathbf{s}_{modele}(\mathbf{c}) = \mathbf{s}_m + \mathbf{Q}_s \mathbf{c} \qquad \mathbf{g}_{modele}(\mathbf{c}) = \mathbf{g}_m + \mathbf{Q}_g \mathbf{c} \tag{2}$$

La procédure d'adaptation de type *descente de gradient* (permettant d'adapter automatiquement le modèle d'appa-

rence à un visage cible) se base sur une recherche itérative du pas optimal à appliquer sur les paramètres d'une certaine configuration de pose et de vecteur $\mathbf{c}$ afin de minimiser le résidu $\mathbf{r}(\mathbf{q})$ entre la texture extraite à cette configuration dans l'image et celle du modèle :

$$\mathbf{r}(\mathbf{q}) = \delta\mathbf{g}(\mathbf{q}) = \mathbf{g}_{image}(\mathbf{q}) - \mathbf{g}_{modele}(\mathbf{q}) \tag{3}$$

$\mathbf{q}$ est le vecteur de paramètres du modèle combiné et/ou des paramètres de pose. L'adaptation automatique du modèle à un visage cible est assurée par un algorithme de type descente de gradient [1]. Le but est de rechercher le $\delta\mathbf{q}$ à appliquer afin de minimiser la norme L_2 du résidu de texture $|r(\mathbf{q}+\delta\mathbf{q})|^2$. Le développement limité de Taylor au premier ordre nous permet d'écrire :

$$r(\mathbf{q}+\delta\mathbf{q}) = \mathbf{r}(\mathbf{q}) + \left(\frac{\partial\mathbf{r}}{\partial\mathbf{q}}\right)\delta\mathbf{q} \tag{4}$$

En mettant l'équation (4) à zéro, on obtient la solution suivante :

$$\delta\mathbf{q} = -\mathbf{R}\mathbf{r}(\mathbf{q}) \tag{5}$$

où la matrice $\mathbf{R}$ est considérée fixe et prédéfinie lors de la construction du modèle d'apparence actif :

$$\mathbf{R} = \left(\left(\frac{\partial\mathbf{r}}{\partial\mathbf{q}}\right)^T\left(\frac{\partial\mathbf{r}}{\partial\mathbf{q}}\right)\right)^{-1}\left(\frac{\partial\mathbf{r}}{\partial\mathbf{q}}\right)^T\mathbf{r}(\mathbf{q}) \tag{6}$$

2.2 Algorithme de Condensation

L'algorithme de Condensation [4] est basé sur les méthodes séquentielles de Monte Carlo connues aussi sous le nom de filtrage particulaire [2]. Il permet d'approcher, à chaque pas temporel t, la distribution de probabilité a posteriori $P(\mathbf{x}_t|\mathbf{z}_{1:t})$ de l'état caché du visage $\mathbf{x}_t$ par une distribution empirique d'un système de particules, où chaque particule est une hypothèse d'état pondérée [5]. Il consiste à propager l'ensemble des particules selon un modèle dynamique $P(\mathbf{x}_t|\mathbf{x}_{t-1})$ et à pondérer chaque particule proportionnellement à sa vraisemblance $P(\mathbf{z}_t|\mathbf{x}_t)$ par rapport aux observations $\mathbf{z}_{1:t}$. Une étape de rééchantillonnage permet de privilégier les particules de poids forts susceptibles de représenter significativement la densité de probabilité a posteriori [4]. Une description plus détaillée de l'algorithme de Condensation est donnée dans ce qui suit :

- A $t = 0$, générer N échantillons $\mathbf{e}_0^{(1)}, \ldots, \mathbf{e}_0^{(N)}$ à partir d'une loi de probabilité initiale $P(\mathbf{x}_0)$ et leur assigner des poids identiques $\pi_0^{(1)} = \ldots = \pi_0^{(N)} = \frac{1}{N}$. C'est l'étape d'*initialisation* du filtre à particules.

- A chaque pas temporel t, on dispose de N particules pondérées $(\mathbf{e}_{t-1}^{(n)}, \pi_{t-1}^{(n)})$, $n = 1, \ldots, N$. Il s'agit alors de :

1. **Rééchantillonner** les particules : tirer N fois les particules avec des probabilités proportionnelles à leurs poids, ceci permet de garder uniquement les particules de poids forts.
2. **Prédire** les N nouvelles particules en échantillonnant à partir du modèle dynamique $P(\mathbf{x}_t|\mathbf{x}_{t-1} = \mathbf{e}_{t-1}^{(n)})$. C'est l'étape de *prédiction* du filtre.
3. **Pondérer** les particules proportionnellement à leur vraisemblance :
$$\pi_t^{(n)} = \frac{P(\mathbf{z}_t|\mathbf{x}_t=\mathbf{e}_t^{(n)})}{\sum_{n=1}^{N} P(\mathbf{z}_t|\mathbf{x}_t=\mathbf{e}_t^{(n))}}$$
l'ensemble des particules pondérées représente une approximation de la densité de probabilité a posteriori. C'est l'étape de *mise à jour* du filtre.
4. **Estimer** l'état optimal $\hat{\mathbf{x}}_t$ par maximisation de la vraisemblance (MAP) :
$$\hat{\mathbf{x}}_t = argmax_{(x_t)}[P(\mathbf{x}_t|\mathbf{z}_{1:t})] \approx argmax_{(e_t^{(n)})}[\pi_t^{(n)}]$$

3 Approche proposée

Comme notre but consiste à suivre les variations de pose et d'apparence d'un visage, nous considérons un vecteur d'état contenant les quatre paramètres de pose 2D $\mathbf{p}_t$ ainsi que les paramètres d'apparence $\mathbf{c}_t$. Selon notre expérimentation, les quatre premières composantes du vecteur $\mathbf{c}_t$ sont capables de représenter la majeur-partie des variations d'apparence (95% de la variance). Le vecteur d'état à l'instant t, noté $\mathbf{x}_t = [\mathbf{p}_t, \mathbf{c}_t]^T$ est alors de dimension 8.

3.1 Modèle dynamique

Nous adoptons un modèle dynamique adaptatif $P(\mathbf{x}_t|\mathbf{x}_{t-1})$ en s'inspirant des idées proposées par Zhou et al [6]. Le modèle a la forme suivante:

$$\mathbf{x}_t = \hat{\mathbf{x}}_{t-1} + \mathbf{v}_t + \mathbf{S}_t\mathbf{u} = \tilde{\mathbf{x}}_t + \mathbf{S}_t\mathbf{u} \tag{7}$$

- $\hat{\mathbf{x}}_{t-1}$ est l'estimation du vecteur d'état à l'instant précédent,
- le vecteur $\mathbf{v}_t = (\partial\mathbf{p}, \partial\mathbf{c})^T$ correspond à la correction prédite par rapport à la pose et à l'apparence,
- $\mathbf{u}$ est un bruit gaussien de moyenne nulle et de variance unitaire,
- la matrice diagonale $\mathbf{S}_t = diag(\sigma_t^{(t_x)}, \ldots, \sigma_t^{(c_4)})$ contient les écarts-types des paramètres de pose et d'apparence.

La correction prédite $\mathbf{v}_t$ est obtenue par adaptation automatique du modèle d'apparence selon une méthode itérative de descente de gradient minimisant un critère résiduel entre la texture modèle et la texture image (§ 2.1) [1]. Nous considérons des écarts-types calculés comme suit :

$$[\sigma_t^{(t_x)}, \ldots, \sigma_t^{(c_4)}]^T = R_t[\sigma_0^{(t_x)}, \ldots, \sigma_0^{(c_4)}]^T \tag{8}$$

avec

$$R_t = diag(R_t^{(t_x)}, \ldots, R_t^{(c_4)})$$

$\sigma_0^{(t_x)}, \ldots, \sigma_0^{(c_4)}$ sont des écarts-types fixes préalablement appris. Les facteurs $R_t^{(i)}$ associés à chacune des 8 composantes du vecteur d'état sont proportionnels à la valeur $\sqrt{\varepsilon_t}$ et appartiennent chacun à un intervalle $[R_{min}^{(i)}, R_{max}^{(i)}]$:

$$R_t^{(i)} = max(min(\sqrt{\varepsilon_t}, R_{max}^{(i)}), R_{min}^{(i)}) \tag{9}$$

où ε_t est une mesure de variance correspondant à une erreur de texture moyennée à travers les L pixels des textures :

$$\varepsilon_t = \frac{2}{L}\sum_{l=1}^{L}\rho\left(\frac{g_{modele}^l(\tilde{\mathbf{c}}_t) - g_{image}^l(\tilde{\mathbf{p}}_t, \tilde{\mathbf{c}}_t)}{\sigma_l}\right) \tag{10}$$

Lorsque les facteurs $R_t^{(i)}$ sont importants, les variances de la distribution prédite et par conséquent l'espace d'état à explorer le sont également, et dans ce cas un nombre important de particules est nécessaire. Nous utilisons donc un nombre de particules adaptatif N_t, obtenu selon l'équation suivante :

$$N_t = N_0 \frac{1}{8} \sum_{i=1}^{8} R_t^{(i)} \quad (11)$$

où N_0 est un nombre de particules fixe prédéfini.

3.2 Modèle d'observation basé sur le modèle d'apparence actif

Le modèle d'observation permet d'évaluer la vraisemblance en chaque particule. Cette vraisemblance est estimée en comparant :

- la texture image échantillonnée à la pose et forme données par l'hypothèse d'état, $\mathbf{g}_{image}(\mathbf{p}_t, \mathbf{c}_t)$ (L'espace d'état caché code les paramètres de pose $\mathbf{p}_t$ et d'apparence $\mathbf{c}_t$),
- et la texture modèle $\mathbf{g}_{modele}(\mathbf{c}_t)$ donnée par le modèle d'apparence et décrite par l'équation (2).

La fonction de vraisemblance que nous avons adoptée a la forme suivante :

$$P(\mathbf{z}_t|\mathbf{x}_t) = Ce^{-d[\mathbf{g}_{modele};\mathbf{g}_{image}]} \quad (12)$$

C est une constante de normalisation de cette distribution. La distance de texture $d[;]$ est proportionnelle à une fonction d'erreur $\rho()$ sommée sur les L pixels de texture et pondérée par la déviation standard σ_l en chaque pixel :

$$d[\mathbf{g};\mathbf{g}'] = \sum_{l=1}^{L} \rho\left(\frac{\mathbf{g}_l - \mathbf{g}'_l}{\sigma_l}\right) \quad (13)$$

Nous avons choisi une fonction d'erreur $\rho()$ robuste ayant la forme suivante [3] :

$$\rho(\lambda) = \begin{cases} \lambda^2 & \text{si } |\lambda| \leq h \\ h|\lambda| - \frac{1}{2}h^2 & \text{si } |\lambda| > h \end{cases} \quad (14)$$

où h est un seuil prédéfini à partir duquel l'erreur $|\lambda|$ est considérée comme aberrante.

3.3 Modèle d'observation calculé et mis à jour à la volée

L'efficacité du modèle d'apparence largement prouvée en littérature reste conditionnée par le fait que l'apparence à suivre doit être préalablement apprise et modélisée. Cette modélisation est donc sensible aux conditions d'enregistrement des images d'apprentissage. Afin de remédier à ce problème, nous remplaons le modèle d'apparence actif (noté $\mathbf{g}_{modele}$ dans la section précédente) par une apparence de texture adaptative calculée *à la volée*, $\mathbf{g}_{volee}$. Ce nouveau modèle présente une robustesse face aux variations d'illumination, et le suivi est adapté à chaque visage sans être conditionné par un apprentissage préalable de son apparence. Ce modèle $\mathbf{g}_{volee}$, initialisé manuellement à l'aide de la texture du visage dans la première image de la séquence vidéo, est actualisé à chaque instant t à l'aide de l'équation suivante :

$$\mathbf{g}_{volee}(t) = \alpha \mathbf{g}_{volee}(t-1) + (1-\alpha)\mathbf{g}_{image}(t, \hat{x}_{t-1}) \quad (15)$$

où α est un facteur d'oubli déterminant l'importance de la mise à jour de la texture modèle. $\mathbf{g}_{image}(t, \hat{x}_{t-1})$ est la texture image courante estimée d'après l'hypothèse d'état $\hat{x}_{t-1}$ retenue à $t-1$. L'espace d'état caché code les paramètres de pose $\mathbf{p}_t$ et de forme $\mathbf{s}_t$ du modèle de visage. La forme $\mathbf{s}_t$ est ici toujours apprise sur une base de visages expressifs (§ 2.1).

4 Résultats expérimentaux

La méthode proposée a été implémentée en C++ et testée sur un PC opérant sous WinXP à 2.4 GHz avec 512 Mb de RAM. Les résultats présentés sont obtenus sur des séquences vidéo contenant un visage vu de face, présentant de larges variations de pose et de mouvements faciaux. Nous avons obtenu des résultas encourageants sur de longues séquences en utilisant le modèle d'observation basé sur le modèle d'apparence actif (FIG. 1).

FIG. 1: Suivi de la pose et des mouvements faciaux, images 085, 366, 615 et 651. Sur chaque image, la forme dessinée correspond à l'état estimé du visage; les textures modèle et image $g_{modele}(\mathbf{c}_t)$ et $g_{image}(\mathbf{p}_t, \mathbf{c}_t)$ sont présentées dans le coin en bas à droite. h = 0.006

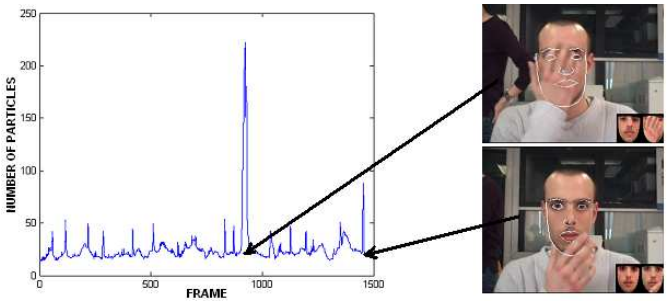

FIG. 2: Evolution du nombre de particules N_t à travers le temps dans une séquence vidéo présentant des occultations du visage. Une occultation importante du visage à l'image 921 provoque un pic. Le pic provoqué par une occultation partielle à l'image 1400 est cependant moins important.

Pour $N_0 = 500$, le nombre de particules N_t évolue entre 20 et 80, et augmente jusqu'à 220 en cas d'occultation du visage (FIG. 2); le temps de traitement est de 2 images par seconde. La figure 3 illustre le résultat du suivi de l'apparence d'un visage, à partir du modèle calculé *à la volée*. Le facteur d'oubli α détermine l'importance de la mise à jour de la texture modèle $\mathbf{g}_{volee}$ par la texture image courante estimée d'après l'hypothèse d'état retenue à l'instant précédent. Afin d'éviter la divergence du suivi dans le cas d'occultation du visage, la valeur du facteur α doit être bornée. Selon notre expérimentation, $\alpha \in [0, 0.25]$.

FIG. 3: Suivi de la pose et des mouvements faciaux utilisant une apparence faciale calculée *à la volée* ; images 75, 215, 470 et 1310 extraites d'une séquence vidéo. $\alpha = 0.2$, h = 0.006

Le suivi de la pose et de l'apparence du visage utilisant le modèle calculé *à la volée* a été testé dans une séquence vidéo présentant des occultations du visage. Les résultats obtenus sont encouragents (FIG. 4) dans le sens où le suivi converge vers la bonne configuration du visage lorsque ce dernier est à nouveau visible dans le support image.

FIG. 4: Suivi de la pose et des mouvements faciaux utilisant le modèle calculé *à la volée* dans une séquence vidéo présentant des occultations du visage; images 52, 54, 56 et 57 ; $\alpha = 0.01$, h = 0.006.

Le tableau suivant illustre les valeurs numériques de certains paramètres utilisés pour les deux approches proposées :

	σ_0	R_{min}	R_{max}
t_x	10	1	50
t_y	10	1	50
s	0.1	0.01	2
θ	1	-2	2

FIG. 5: valeurs des écarts types initiaux σ_0 ainsi que des facteurs R_{min} et R_{max}, relatifs à la position en x (t_x) et y (t_y), au facteur d'échelle s et à l'angle de rotation θ (radian).

5 Conclusion et perspectives

Nous proposons dans cet article une approche stochastique permettant de suivre les variations de pose et d'apparence d'un visage quasi frontal dans les séquences vidéo. Cette approche repose sur le principe de filtrage particulaire connu sous le nom d'algorithme de Condensation dans le domaine de vision par ordinateur. La dynamique des particules est adaptative, guidée par une optimisation déterministe. La distribution des observations est dérivée d'un modèle d'apparence actif et estimée en intégrant des mesures robustes. Des résultats sont présentés, notamment en prenant en compte des phases d'occultation du visage. En second lieu nous introduisons une deuxième approche qui consiste à remplacer le modèle d'apparence actif par une apparence de texture mise à jour à la volée (en ligne) afin de remédier à la contrainte de l'apprentissage préalable de l'apparence à suivre et au problème de sensibilité du modèle d'apparence aux conditions d'enregistrement des images d'apprentissage. Les résultats expérimentaux sont également encourageants. Une extension intéressante à ces travaux concerne la reconnaissance de l'évolution des expressions faciales une fois le visage et ses mouvements faciaux déterminés.

References

[1] T. F. Cootes, G. J. Edwards et C. J. Taylor. *Active Appearance Models*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 23(6) pp. 681-685, juin 2001.

[2] A. Doucet, J. F. G. De Freitas et N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[3] P. J. Huber. *Robust Statistics*. Wiley, 1981.

[4] M. Isard et A. Blake. *Condensation Conditional Density Propagation for Visual Tracking*. Int. Journal of Computer Vision, pp. 5-28, 1998.

[5] J. McCormick. *Stochastic Algorithms for Visual Tracking*. Springer-Verlag, 2002.

[6] S. Zhou, R. Chellappa, et B. Moghaddam. *Visual tracking and recognition using appearance-adaptive models in particle filters*. IEEE Trans. on Image Processing, 13(11), pp. 1491-1506, 2004.