

Suivi des variations de pose et d'apparence de visages dans des séquences vidéo

S. Hamlaoui M. Dang F. Davoine

Laboratoire HEUDIASYC - CNRS / UTC

Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex - France

{shamlaou, mdang, fdavoine}@hds.utc.fr

Résumé

Nous proposons dans cet article un système stochastique de suivi de visages vus de face dans des séquences vidéo. Ce système permet de suivre simultanément la pose 2D du visage (position dans le support image, facteur d'échelle, angle de rotation) ainsi que les gestes faciaux internes (expressions faciales) représentés par les variations d'apparence du visage, l'apparence étant décrite par des paramètres contrôlant la forme et la texture du visage. Notre système se base essentiellement sur le principe du filtrage particulaire et plus précisément sur l'algorithme de Condensation. Nous utilisons également le modèle d'apparence actif (AAM) développé par Cootes et Taylor pour modéliser les variations d'apparence du visage. Nous présentons des résultats de suivi obtenus par une approche déterministe et ceux obtenus par le suivi stochastique représenté par notre système.

Mots clefs

Filtrage particulaire, algorithme de Condensation, suivi stochastique de visages, modèles d'apparence actifs.

1 Introduction

La modélisation des systèmes dynamiques et l'analyse du mouvement font l'objet d'un grand intérêt dans la communauté de la vision par ordinateur. Notre étude se focalise sur le suivi de visages dans les séquences vidéo, et plus précisément sur le suivi des paramètres de pose et des variations d'apparence interne : une problématique à laquelle on est confronté dans plusieurs applications telles que la télésurveillance, les interfaces homme-machine, la détection d'hypovigilance des conducteurs, les systèmes de téléconférence, etc.

Dans la littérature, les méthodes de suivi rencontrées se décomposent généralement en deux grandes catégories : déterministes et stochastiques.

Les méthodes déterministes consistent à rechercher les paramètres de mouvement optimaux permettant de minimiser l'erreur entre un modèle de l'objet suivi et l'information extraite à chaque image de la vidéo. Elles sont donc généralement réduites à un problème

d'optimisation [1]. L'optimisation se base alors sur une recherche déterministe telle que dans les méthodes classiques de descente de gradient, etc. Le problème souvent rencontré dans ces méthodes est la présence de minima locaux qui peuvent provoquer la divergence de la recherche par rapport au minimum global visé.

Les méthodes stochastiques, quant à elles, se basent sur une recherche probabiliste de l'hypothèse de configuration la plus vraisemblable par rapport à un modèle prédéfini de l'objet suivi. Ces méthodes se basent essentiellement sur le principe du filtrage Bayésien [2], où le but est d'estimer un état caché représentant les paramètres du mouvement au vu d'une série d'observations bruitées des états passés du système, correspondant aux données images. L'évolution des états à travers le temps est décrite par un modèle dynamique Markovien. Un modèle d'observation permet d'estimer la vraisemblance de chaque hypothèse d'état. Le suivi correspond alors à un problème d'estimation récursive de la densité de probabilité d'état conditionnellement aux observations à travers les étapes de prédiction et de mise à jours du filtre.

L'implémentation du filtrage Bayésien est basée sur l'estimation séquentielle de Monte Carlo [3] connue aussi sous le nom de filtrage particulaire [4]. Cette approche est connue pour sa robustesse aux non-linéarités du système et aux non-Gaussianités des lois de probabilités manipulées. Le filtrage particulaire est une méthode d'approximation, permettant de représenter la densité d'état a posteriori par un ensemble aléatoire d'échantillons pondérés (particules), où chaque particule est une hypothèse d'état pondérée [5].

L'une des méthodes les plus connues se basant sur le principe du filtrage particulaire et appliquées au domaine de vision par ordinateur est l'algorithme de Condensation proposé par Isard et Blake [6, 7].

Le système que nous proposons dans cet article permet de suivre simultanément la pose 2D du visage ainsi que les gestes faciaux internes représentés par les variations d'apparence du visage. Ce système se base essentiellement sur le principe de l'algorithme de Condensation. Nous utilisons aussi le modèle d'apparence actif (AAM) développé par Cootes et Taylor comme modèle du visage [8, 9]. L'AAM combine deux modèles

statistiques décrivant les variations de forme et de texture et est présenté dans la première section de cet article. Dans la deuxième section, nous décrivons l'approche du suivi déterministe par AAM. Notre approche est détaillée dans la section 3 où une définition plus explicite de l'algorithme de Condensation est rappelée.

Nous présentons dans la section 4 les résultats de suivi obtenus par notre système et ceux obtenus par le suivi déterministe. Ces résultats sont alors discutés par la suite. La cinquième section est consacrée à la conclusion et la présentation de perspectives.

2 Modèle d'Apparence Actif

Le modèle d'apparence actif [8] est une représentation statistique linéaire des variations d'apparence de la classe de visages par une Analyse en Composantes Principales. Les paramètres de ce modèle contrôlent les variations de forme et de texture et sont préalablement appris à partir d'un ensemble d'images annotées de façon à définir les structures faciales. Les formes sont définies par les coordonnées spatiales des points d'annotation. Les textures quant à elles correspondent à l'intensité des pixels à l'intérieur de la région délimitée par chacune des formes. Ces formes sont d'abord alignées à une forme moyenne par le biais d'une transformation Procrustéenne généralisée [8]. Une ACP est par la suite appliquée à l'ensemble des formes \mathbf{s} et des textures \mathbf{g} :

$$\mathbf{s} = \mathbf{s}_m + \boldsymbol{\phi}_s \mathbf{b}_s \quad (1)$$

$$\mathbf{g} = \mathbf{g}_m + \boldsymbol{\phi}_g \mathbf{b}_g \quad (2)$$

\mathbf{b}_s , \mathbf{b}_g sont respectivement les paramètres du modèle de forme et de texture. \mathbf{s}_m et \mathbf{g}_m représentent la forme et la texture moyennes et $\boldsymbol{\phi}_s$, $\boldsymbol{\phi}_g$ sont les vecteurs propres des matrices de covariance de forme et texture.

Afin d'obtenir un modèle combiné de forme et texture, les paramètres des deux modèles sont alors couplés de la manière suivante :

$$\mathbf{b} = \begin{bmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{bmatrix} \quad (3)$$

\mathbf{W}_s étant une matrice de normalisation. Une troisième ACP est alors appliquée afin d'obtenir le vecteur \mathbf{c} du modèle d'apparence combiné :

$$\mathbf{b} = \boldsymbol{\phi}_c \mathbf{c} \quad (4)$$

où l'ensemble des vecteurs propres du modèle combiné est représenté par :

$$\boldsymbol{\phi}_c = \begin{bmatrix} \boldsymbol{\phi}_{c,s} \\ \boldsymbol{\phi}_{c,g} \end{bmatrix}. \quad (5)$$

De nouvelles instances de forme et texture $\mathbf{s}_{\text{modèle}}$ et $\mathbf{g}_{\text{modèle}}$ peuvent alors être générées à partir du vecteur \mathbf{c} :

$$\mathbf{s}_{\text{modèle}}(\mathbf{c}) = \mathbf{s}_m + \boldsymbol{\phi}_s \mathbf{W}_s^{-1} \boldsymbol{\phi}_{c,s} \mathbf{c} = \mathbf{s}_m + \mathbf{Q}_s \mathbf{c} \quad (6)$$

$$\mathbf{g}_{\text{modèle}}(\mathbf{c}) = \mathbf{g}_m + \boldsymbol{\phi}_g \boldsymbol{\phi}_{c,g} \mathbf{c} = \mathbf{g}_m + \mathbf{Q}_g \mathbf{c}. \quad (7)$$

3 Approche déterministe

Le suivi déterministe que nous présentons est l'une des applications des modèles d'apparence actifs. Le suivi déterministe par AAM 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 $\mathbf{r}(\mathbf{p})$ entre la texture extraite à cette configuration dans l'image et celle du modèle :

$$\mathbf{r}(\mathbf{p}) = \delta \mathbf{g}(\mathbf{p}) = \mathbf{g}_{\text{image}}(\mathbf{p}) - \mathbf{g}_{\text{modèle}}(\mathbf{p}) \quad (8)$$

\mathbf{p} 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 [8]. Le but est de rechercher le $\delta \mathbf{p}$ à appliquer afin de minimiser la norme L_2 du résidu de texture $|\mathbf{r}(\mathbf{p} + \delta \mathbf{p})|^2$. Le développement limité de Taylor au premier ordre nous permet d'écrire :

$$\mathbf{r}(\mathbf{p} + \delta \mathbf{p}) = \mathbf{r}(\mathbf{p}) + (\partial \mathbf{r} / \partial \mathbf{p}) \delta \mathbf{p}. \quad (9)$$

La solution est alors du type :

$$\delta \mathbf{p} = ((\partial \mathbf{r} / \partial \mathbf{p})^T (\partial \mathbf{r} / \partial \mathbf{p}))^{-1} (\partial \mathbf{r} / \partial \mathbf{p})^T \mathbf{r}(\mathbf{p}) = -\mathbf{R} \mathbf{r}(\mathbf{p}). \quad (10)$$

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

4 Approche stochastique proposée

Dans la problématique de suivi stochastique, il s'agit d'estimer à chaque instant t les paramètres de mouvement représentés par un vecteur d'état caché \mathbf{x}_t , de dimension k au vu d'une série d'observations bruitées des états passés du système $\mathbf{z}_{1:t}$, représentées par des données image à travers un vecteur d'observation \mathbf{z}_t de dimension M . La transition d'un état à l'autre à travers le temps est décrite par un modèle dynamique Markovien $P(\mathbf{x}_t | \mathbf{x}_{t-1})$. La vraisemblance de chaque hypothèse d'état est estimée à partir d'un modèle d'observation $P(\mathbf{z}_t | \mathbf{x}_t)$. Ces deux modèles sont supposés connus. Il s'agit alors d'évaluer récursivement la densité de probabilité a posteriori de l'état conditionnellement aux observations $P_t(\mathbf{x}_t | \mathbf{z}_{1:t})$.

Comme notre but consiste à suivre les variations de pose et d'apparence du visage, nous considérons que le vecteur d'état contient les paramètres d'apparence \mathbf{c}_t ainsi que les quatre paramètres de pose $\mathbf{pose}_t = (t_x, t_y, sf, \theta)_t$, représentant la position 2D, le facteur d'échelle et l'angle de rotation de la forme du visage suivi :

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{pose}_t \\ \mathbf{c}_t \end{bmatrix}. \quad (11)$$

Le vecteur d'observation \mathbf{z}_t représente la texture extraite à chaque image et décrite par l'intensité des pixels :

$$\mathbf{z}_t = \mathbf{g}_{\text{image}}. \quad (12)$$

4.1 Algorithme de Condensation

Notre approche se base essentiellement sur l'algorithme de Condensation [6, 7] qui consiste à :

1- Générer N échantillons $\mathbf{e}_0^{(1)}, \dots, \mathbf{e}_0^{(N)}$ à partir d'une loi de probabilité initiale $P(\mathbf{x}_0)$ et leur assigner des poids identiques $\pi_0^{(1)}, \dots, \pi_0^{(N)}$. C'est l'étape d'initialisation du filtre.

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

- i. **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.
- ii. **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.
- iii. **Pondérer** les particules proportionnellement à leur vraisemblance :

$$\pi_t^{(n)} = P(\mathbf{z}_t | \mathbf{x}_t = \mathbf{e}_t^{(n)}) / \sum_{n=1}^N P(\mathbf{z}_t | \mathbf{x}_t = \mathbf{e}_t^{(n)}) \quad (13)$$

Ces poids représentent une approximation de l'amplitude de la densité de probabilité a posteriori. C'est l'étape de mise à jour du filtre.

- iv. **Estimer** l'état optimal par maximisation de la vraisemblance (MAP) :

$$\text{MAP: } \underset{\mathbf{x}_t}{\operatorname{argmax}} P(\mathbf{x}_t | \mathbf{z}_{1:t}) \approx \underset{\mathbf{s}_t^{(n)}}{\operatorname{argmax}} \pi_t^{(n)} \quad (14)$$

ou par calcul de la moyenne

$$E(\mathbf{x}_t | \mathbf{z}_{1:t}) \approx \sum_{n=1}^N \mathbf{e}_t^{(n)} \pi_t^{(n)}. \quad (15)$$

4.2 Modèle dynamique

Nous avons adopté un modèle dynamique à vitesse nulle plutôt que constante afin de mieux gérer les changements brusques de vitesse et de direction du mouvement apparent. Ce modèle est décrit par la densité Gaussienne suivante :

$$P(\mathbf{x}_t | \mathbf{x}_{t-1}) = G(\mathbf{x}_t | \mathbf{x}_{t-1}, \Sigma_x) \quad (16)$$

La moyenne de la Gaussienne est estimée à l'état précédent. La matrice de covariance Σ_x est supposée diagonale dans un but de simplicité :

$$\Sigma_x = \operatorname{diag}(\sigma_{x1}^2, \dots, \sigma_{xk}^2) \quad (17)$$

Cette hypothèse implique que les paramètres de pose et d'apparence sont supposés être mutuellement indépendants. Leurs variances σ_{xj}^2 sont estimées à partir d'un ensemble d'apprentissage.

4.3 Modèle d'observation

Le modèle d'observation permet d'évaluer la vraisemblance en chaque particule. Cette vraisemblance est estimée en comparant la donnée image décrite par l'observation \mathbf{z}_t et la donnée déterminée par l'hypothèse d'état \mathbf{x}_t . La fonction de vraisemblance que nous avons adoptée a la forme suivante :

$$P(\mathbf{z}_t | \mathbf{x}_t) = C \exp \left[-\frac{1}{2} D_m(\mathbf{g}_{\text{modèle}}(\mathbf{c}_t), \mathbf{g}_{\text{image}}) \right], \quad (18)$$

C étant une constante de normalisation de la densité. $D_m(\mathbf{g}_{\text{modèle}}(\mathbf{c}_t), \mathbf{g}_{\text{image}})$ correspond à la distance de Mahalanobis entre la texture image et la texture modèle. Cette distance est estimée par l'équation suivante :

$$D_m(\mathbf{g}_{\text{modèle}}(\mathbf{c}_t), \mathbf{g}_{\text{image}}) = \sum_{i=1}^M [(\mathbf{g}_{\text{modèle},i} - \mathbf{g}_{\text{image},i})^2 / \sigma_{gi}^2] \quad (19)$$

M est le nombre de pixels à partir desquels la texture est extraite. $\mathbf{g}_{\text{modèle},i}$, $\mathbf{g}_{\text{image},i}$ correspondent respectivement aux textures modèle et image du pixel i . σ_{gi}^2 représente la variance de texture du pixel i .

La texture modèle $\mathbf{g}_{\text{modèle}}(\mathbf{c}_t)$ est décrite par l'équation (7) et la texture $\mathbf{g}_{\text{image}}$ est extraite de l'image tel que :

$$\mathbf{g}_{\text{image}} = W(\mathbf{z}_t | A(\mathbf{s}_{\text{modèle}}(\mathbf{c}_t), \mathbf{pose}_t) \rightarrow \mathbf{s}_{\text{ref}}). \quad (20)$$

Donc tout d'abord, l'instance de forme $\mathbf{s}_{\text{modèle}}(\mathbf{c}_t)$ générée à partir du modèle est ramené à l'hypothèse de pose \mathbf{pose}_t par une transformation affine A . Un warping W est ensuite appliqué afin de l'aligner à une forme de référence \mathbf{s}_{ref} permettant ainsi d'en extraire la texture image.

5 Résultats expérimentaux

Dans les deux approches déterministe et stochastique, le but est de minimiser le résidu de texture en comparant la donnée image à un modèle d'apparence actif prédéfini et connu. Nous avons testé ces deux approches sur des vidéos représentant des visages expressifs de pose et d'apparence variables et construits à partir de la base de donnée CMU [10]. Les résultats obtenus par l'approche que nous avons proposée sont présentés dans la figure 1. La figure 2 présente le résultat du suivi déterministe par AAM sur deux images successives de la vidéo où les paramètres de pose changent brusquement.

Pour notre implémentation, nous avons utilisé 500 particules. Le vecteur d'état, de dimension 8, contient les quatre paramètres de pose ainsi que les paramètres des quatre premiers modes du modèle d'apparence combiné. L'état optimal maximisant la densité a posteriori est exposé sur les images de test et représente l'hypothèse de pose et de forme donnée par la particule ayant le poids maximal.

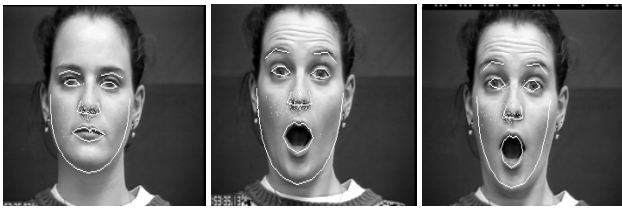


Figure 1 - Suivi des variations d'apparence et de pose par l'approche proposée (images 2-14-15)



Figure 2 - Suivi déterministe des variations d'apparence et de pose (images 14-15)

Nous remarquons dans la figure 2 que lorsque la pose change brusquement d'une image à l'autre, le suivi déterministe semble converger vers un minimum local et ne retrouve donc pas la configuration optimale. Ce problème n'apparaît pas dans le cas du suivi stochastique car les particules représentant les multiples hypothèses de configurations couvrent une plus grande partie de l'espace d'état.

6 Conclusion et perspectives

Nous avons présenté dans cet article une implémentation d'un système stochastique permettant de suivre simultanément les variations de pose et d'apparence des visages.

Cependant, plusieurs directions doivent encore être explorées. Nous citons par exemple l'étude de la possibilité de réduire le nombre des particules ou de le rendre interactif en l'adaptant aux besoins effectifs du suivi à chaque pas temporel, ceci permettra de gagner en temps de calcul du système. L'efficacité de ce système peut aussi être améliorée en intégrant la méthode d'échantillonnage d'importance [11]. Cette méthode utilise une information additionnelle permettant de concentrer une partie des particules dans les régions d'espace d'état ayant une forte probabilité de contenir l'état optimal. Cette information additionnelle peut correspondre par exemple à des attributs de mouvement, de couleur, etc.

La performance de notre approche doit aussi être testée dans le cas d'occlusion. Nous testerons alors l'approche des distances robustes [12]. Dans le cas d'occlusions globales, l'échantillonnage d'importance nous permet aussi de réinitialiser le filtre et par conséquent le suivi en générant des particules dans les régions de l'espace d'état

contenant le plus d'informations sur la configuration a posteriori des états.

Références

- [1] S. Baker, I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision*, 56 (3): 221 - 255, February 2004.
- [2] J. Vermaak, N.D. Lawrence, P. Pérez. Variational Inference for Visual Tracking. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Madison, U.S.A, 2003.
- [3] A. Doucet, J.F.G De Freitas, N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.
- [4] S. Arulampalam, S. Maskell, N. Gordon and T. Clapp. A Tutorial on Particle Filters for On-line Non-Linear / Non-Gaussian Bayesian Tracking. *IEEE Transactions of Signal Processing*, 50(2): 174-188, Février 2002.
- [5] J. McCormick. *Stochastic Algorithms for Visual Tracking*. Springer-Verlag, 2002.
- [6] A. Blake, M. Isard. *Active Contours*. Springer-Verlag, 1998.
- [7] M. Isard, A. Blake. Condensation – Conditional Density Propagation for Visual Tracking. *Int. Journal of Computer Vision*, 29(1) : 5-28, 1998.
- [8] T.F. Cootes, G.J. Edwards and C.J. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6): 681-685, Juin 2001.
- [9] B. Abboud, F. Davoine, M. Dang, Statistical Modelling for Facial Expression Analysis and Synthesis. *IEEE ICIP*. Pages 14-17, Espagne, Septembre 2003.
- [10] T. Kanade, J. Cohn, and Y.L. Tian. Comprehensive Database for Facial Expression Analysis. *In. Proc. of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*. Pages 46-53, Grenoble, France, Mars 2000.
- [11] M. Isard, A. Blake. Icondensation - Unifying Low-Level and High Level Tracking in a Stochastic Framework. *In. Proc. of European Conference on Computer Vision*, vol. 1, pages 893-908, 1998.
- [12] P.J. Huber. *Robust Statistics*. Wiley, 1981.