Publié au GRETSI 2003, Paris, France.

Modélisation implicite du mouvement en suivi par filtrage de Monte Carlo séquentiel

Jean-Marc ODOBEZ[1], Sileye BA[1 *]

[1]IDIAP, Rue du Simplon 4
Case postale 592 CH 1920 Martigny, SUISSE

Jean-Marc.Odobez@idiap.ch, Sileye.Ba@idiap.ch

Résumé – Le filtrage par méthode de Monte-Carlo séquentiel (MCS) est l'une des méthodes les plus populaires pour effectuer du suivi visuel. Dans ce contexte, il est généralement fait l'hypothèse que, étant donnée la position d'un objet dans des images successives, les observations extraites des images de cet objet sont indépendantes. Dans cet article, nous soutenons que, au contraire, ces observation sont fortement corrélées. Pour prendre en compte cette correlation, nous proposons un nouveau modèle qui peut s'interpréter comme l'ajout d'un terme de vraisemblance modélisant implicitement des mesures de mouvement. Le nouveau modèle permet de lever des ambiguïtés visuelles tout en gardant des modèles d'objet simples, comme le montrent les résultats obtenus sur plusieurs séquences et modèles d'objets différents (contour ou distribution de couleurs).

Abstract – Particle filters are now established as the most popular method for visual tracking. Within this framework, it is generally assumed that the data are temporally independent given the sequence of object states. In this paper, we argue that in general the data are correlated, and that modeling such dependency should improve tracking robustness. To take data correlation into account, we propose a new model which can be interpreted as introducing a likelihood on implicit motion measurements. The proposed model allows to filter out visual distractors when tracking objects with generic models based on shape or color distribution representations, as shown by the reported experiments.

1 Introduction

Le suivi d'objets est un problème important en vision par ordinateur. Néanmoins, bien qu'étant étudié de façon intensive, cela reste un problème difficile en présence d'ambiguïtés (e.g. lors du suivi d'un objet en présence d'objets de cette même classe), de bruit dans les mesures (e.g. les problèmes d'illumination), de variabilité de la classe d'objets considérée.

La mise en œuvre d'un algorithme de suivi nécessite la définition de deux éléments principaux : la représentation de l'objet et sa dynamique. La représentation de l'objet correspond à tout ce qui, implicitement ou explicitement, caractérise l'objet : sa position, son apparence, son mouvement etc. Par exemple, des modèles de contour paramétrisés [3] ou des distributions de couleur [1, 4] sont souvent utilisés. Un inconvénients de ces représentations est qu'elles sont peu spécifiques, ce qui les rend sensibles aux ambiguïtés locales. Une manière de rendre ces modèles uniques est d'utiliser des prototypes (template) [5, 6], ce qui conduit à des algorithmes plus robustes. L'inconvénient de ce type d'approches est de n'autoriser que de faibles variations d'apparence dans la séquence, à moins de procéder à une adaptation difficile du modèle ou d'employer des modèles d'apparence plus complexes (e.g. par espace propre).

La dynamique de l'objet est souvent utilisée pour prédire l'espace de recherche de la nouvelle positions de l'objet en fonction de son passé. La difficulté de modélisation de ce terme provient de deux aspects contradictoires. D'un coté, on souhaite que l'espace de recherche soit suffisament grand pour pouvoir appréhender des changement brutaux de mouvement. De l'autre, on souhaite le restreindre pour éviter au suivi d'être perturbé par des ambiguïtés locales proches de la véritable configuration de l'objet, ce qui est susceptible de se produire lors de l'utilisation de modèles d'objets peu spécifiques.

Dans cet article, nous proposons un algorithme de suivi basé sur le filtrage de Monte Carlo Sequentiel (MCS), qui a prouvé son intérêt dans ce domaine [2, 3, 5]. Plus précisément, nous montrerons que l'une des hypothèses standard, l'indépendance des observations conditionnelement à la séquence d'états, n'est pas appropriée dans le cas du suivi visuel. Nous proposons alors un nouveau modèle. D'un point de vue qualitatif, celui-ci permet de modéliser implicitement le mouvement entre deux images. Les avantages du nouveau modèle sont doubles : il permet d'une part de rendre des modèles d'objet peu spécifiques (basés sur des contours, des histogrammes de couleur) moins sensibles aux ambiguïtés; de ce fait, il permet d'autre part d'éviter le dilemme évoqué dans le choix d'un modèle (et des paramètres) de la dynamique d'un objet.

2 Filtrage MCS

Dans cette approche baysienne, la distribution à posteriori $\mathrm{p}(c_{0:k}|I_{0:k})$ de la séquence d'états étant donné les observations est représentée par un ensemble d'échantillons pondérés $\{c^i_{0:k}, w^i_k\}_{i=1}^{N_s}$ (avec $\sum_i w^i_k = 1$) de telle sorte que :

$$\mathrm{p}(c_{0:k}|I_{1:k}) \approx \sum_{i=1}^{N_s} w^i_k \; \delta(c_{0:k} - c^i_{0:k}) \qquad (1)$$

Les poids sont choisis suivant le principe de l'"Importance Sampling" (IS). Plus précisément, si l'on suppose que les échantil-

*Les auteurs remercient le Fond National Suisse de la Recherche Scientifique qui finance ce travail au travers du Pole National de Recherche "Interactive Multimodal Information Management (IM2)".

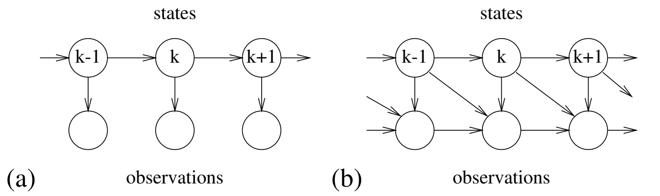


FIG. 1: *Modèle graphique pour le suivi. (a) modèle standard (b) modèle proposé.*

lons sont tirés aléatoirement suivant une fonction de proposition $\mathrm{q}(c_{0:k}|I_{1:k})$, alors les poids conduisant à l'approximation (1) s'expriment par : $w_k^i \propto \frac{\mathrm{p}(c_{0:k}^i|I_{1:k})}{\mathrm{q}(c_{0:k}^i|I_{1:k})}$. Pour déterminer ces poids, on exploite l'équation de recursion suivante sur la distribution a posteriori :

$$\mathrm{p}(c_{0:k}|I_{1:k}) = \frac{\mathrm{p}(I_k|c_{0:k}, I_{1:k-1})\mathrm{p}(c_k|c_{0:k-1}, I_{1:k-1})}{\mathrm{p}(I_k|I_{1:k-1})} \quad (2)$$
$$\times \quad \mathrm{p}(c_{0:k-1}|I_{1:k-1}) \quad (3)$$

et on effectue généralement les hypothèses suivantes :

1. La séquence d'états $c_{0:k}$ suit un modèle markovien d'ordre 1, caractérisé par la définition de la dynamique $\mathrm{p}(c_k|c_{k-1})$.
2. Les observations $\{I_k\}$, conditionellement à la séquence d'états, sont indépendantes. Ceci conduit à $$\mathrm{p}(I_{1:k}|c_{0:k}) = \prod_{i=1}^{k} \mathrm{p}(I_k|c_k),$$ nécessitant la définition des vraisemblances $\mathrm{p}(I_k|c_k)$;
3. La distribution *a priori* $\mathrm{p}(x_{0:k})$ est employée comme fonction de proposition importance. Dans ce cas : $\mathrm{q}(c_{0:k}|I_{1:k}) = \mathrm{q}(c_{0:k}) = \mathrm{p}(c_k|c_{k-1})\mathrm{q}(c_{0:k-1})$

On aboutit alors à l'équation de remise à jour des poids [2] :

$$w_k^i \propto w_{k-1}^i \ \mathrm{p}(I_k|c_k^i) \quad (4)$$

Un rééchantillonnage est nécessaire pour éviter la dégénérescence des échantillons (i.e. éviter que tous les poids sauf un tendent vers zéro) [2]. S'il est effectué à chaque itération, on obtient l'algorithme de bootstrap suivant :

1. Initialisation : pour $i = 1, \ldots, N_s$, échantillonnage suivant $c_0^i \sim \mathrm{p}(c_0)$; poser $k = 1$
2. Etape d' "Importance sampling" : pour $i = 1, \ldots, N_s$, échantillonnage suivant $\tilde{c}_k^i \sim \mathrm{p}(c_k|c_{k-1}^i)$, évaluation des poids suivant $\tilde{w}_k^i \propto w_{k-1}^i \ \mathrm{p}(I_k|\tilde{c}_k^i)$; puis normalisation des poids $\tilde{w}_k^i$.
3. Etape de sélection : Rééchantillonnage de N_s échantillons $\{c_k^i, w_k^i = \frac{1}{N_s}\}$ à partir des échantillons $\{\tilde{c}_k^i, \tilde{w}_k^i\}$; poser $k = k + 1$ et retourner à l'étape 2.

3 Discussion, modèle proposé

L'algorithme présenté ci-dessus repose sur le modèle graphique probabiliste standard présenté à la figure 1a, en accord avec les hypothèses 1 and 2 de la section précédente.

Si la première de ces hypothèses est relativement raisonable, la seconde est en revanche rarement vérifiée en pratique dans le cas du suivi visuel[1]. En effet, dans la plupart des modèles de suivi, la configuration de l'objet inclu des paramètres d'une transformation géométrique $\mathcal{T}$. Celle-ci permet d'extraire explicitement ou implicitement de l'image courante la région associée à l'objet considéré selon :

$$\tilde{I}_{c_t}(\mathbf{r}) = I_t(\mathcal{T}_{c_t}\mathbf{r}), \forall \mathbf{r} \in R, \quad (5)$$

où $\mathbf{r}$ désigne une position, R est une région de référence fixée, et $\mathcal{T}_{c_k}\mathbf{r}$ correspond à l'application de la transformation $\mathcal{T}$ paramétrisée par c_k au pixel $\mathbf{r}$. La vraisemblance des données est alors calculée à partir de ce patch local : $\mathrm{p}(I_k|c_k) = \mathrm{p}(\tilde{I}_{c_k})$. Or, si c_{k-1} et c_k sont deux états successifs d'un objet, on peut faire l'hypothèse suivante :

$$\tilde{I}_{c_k}(\mathbf{r}) = \tilde{I}_{c_{k-1}}(\mathbf{r}) + bruit \qquad \forall \mathbf{r} \in R \quad (6)$$

où $bruit$ prend généralement une valeur très faible. Ce point est illustré sur la figure 2. L'équation (6) est à la base de tous les algorithmes d'estimation et de compensation de mouvement comme MPEG. Ainsi, d'après cette équation, l'indépendance des données conditionellement à la séquence d'états n'est pas valide. Plus précisément :

$$\mathrm{p}(I_k|I_{1:k-1}, c_{1:k}) \neq \mathrm{p}(I_k|c_k). \quad (7)$$

FIG. 2: *Images aux temps t et $t + 3$. Les deux patchs locaux sont fortement corrélés.*

Compte tenu de (6), un modèle plus approprié pour le suivi visuel est représenté sur la figure 1b.

Le nouveau modèle peut être incorporé dans le filtre MCS. Partant de (3), toutes les équations suivantes peuvent être dérivées en utilisant la nouvelle hypothèse. En conservant les hypothèses 1 et 3, cela conduit à l'équation de remise à jour suivante :

$$w_k^i \propto w_{k-1}^i \ \mathrm{p}(I_k|I_{k-1}, c_k^i, c_{k-1}^i) \quad (8)$$

en remplacement de (4) dans l'algorithme de bootstrap.

3.1 Espace d'état, modèle dynamique

Pour représenter l'objet, nous suivons une approche 2D standard, où l'objet est représenté par une région R et son contour. Ce dernier est caractérisé par une forme paramétrique, dans notre cas une ellipse. La transformation géométrique, auquel l'objet est sujet, est composée d'une translation $\mathbf{T}$ et d'un facteur d'échelle s. De plus, nous avons choisi comme modèle dynamique un modèle auto-régressif du premier ordre classique appliqué aux paramètres augmentés de leur dérivée première. Plus précisément, en désignant l'état par $c_k = (\alpha_k, \dot{\alpha}_k)$ avec $\alpha = (\mathbf{T}, s)$, le modèle dynamique s'écrit : $c_k = Ac_{k-1} + B\mathbf{w}_k$ où A et B sont les paramètres (fixes) du modèle et $\mathbf{w}$ est un bruit blanc.

3.2 Vraisemblance des données

Pour appliquer le nouveau modèle, nous considérons la vraisemblance de données suivante :

$$\mathrm{p}(I_k|I_{k-1}, c_k, c_{k-1}) = \mathrm{p}_c(I_k|I_{k-1}, c_k, c_{k-1}) \times \mathrm{p}_o(I_k|c_k) \quad (9)$$

où p_c modélise la correlation temporelle et p_o modélise la vraisemblance de l'objet. Ce choix découple la modélisation de la

1. Dans le cas du suivi de contours, l'hypothèse est relativement valide dans la mesure où la fonction d'autocorrelation temporelle est très pointue.

FIG. 3: *Suivi de la région délimitée par le cadre blanc dans l'image de gauche : les particules associées aux cadres verts devraient avoir une plus grande probabilité que celle associée au cadre rouge.*

correlation existant entre deux images consécutives d'un même objet, dont le but implicite est de s'assurer que la trajectoire de l'objet suit le flux optique, de la modélisation de la forme (ou de l'apparence) de l'objet. Nous avons fait l'hypothèse que ces deux termes étaient indépendants. Lorsque l'objet est modélisé par son contour, cette hypothèse est valide dans la mesure où l'évaluation de la vraisemblance de l'objet fait intervenir des données sur le pourtour de l'objet alors que le terme de corrélation s'applique essentiellement à l'intérieur de l'objet.

Vraisemblance de contour

La vraisemblance d'objet suit le modèle présenté dans [3], où des mesures de contours sont calculées le long de L lignes normales à une ellipse hypothèse supposée située sur un fond bruité. On obtient :

$$\mathrm{p}_o(I_t|c_t) \propto \prod_{l=1}^{L} \max\left(K, \exp(-\frac{\|\hat{\nu}^l_{min} - \nu^l_0\|^2}{2\sigma^2})\right), \qquad (10)$$

où $\hat{\nu}^l_{min}$ représente le point de contour le plus proche du point ν^l_0 (se trouvant sur le contour hypothèse) détecté sur la $l^{\text{ième}}$ ligne, et K est une constante introduite quand aucun contour n'est détecté.

Vraisemblance de correlation

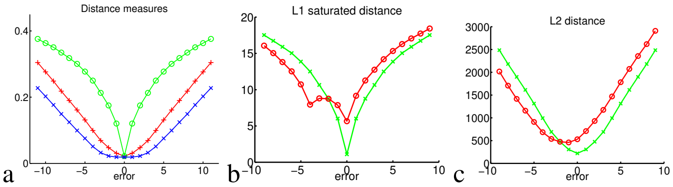
Nous modélisons ce terme par :

$$\mathrm{p}_c(I_k|I_{k-1}, c_k, c_{k-1}) \propto \exp^{-\lambda_c \mathrm{d}_c(\tilde{I}_{c_k}, \tilde{I}_{c_{k-1}})} \qquad (11)$$

où d_c représente une distance entre deux patches. Beaucoup de telles distance ont été définies dans la littérature [5, 7]. Le choix de cette distance doit prendre en compte les considérations suivantes, illustrées par la figure 3.

1. la distance doit modéliser l'information de mouvement sous-jacente, i.e. la distance doit augmenter si l'erreur de prédiction augmente;
2. l'aspect aléatoire du processus de prédiction dans le filtre MCS produit rarement des configurations correspondant au match exact;
3. dans le cas où l'objet et le fond ont des mouvements différents, la distance entre une particule et ses prédictions doit en général être plus faible lorsque cette particule ne recouvre que l'objet (cadre blanc de la figure 3) que lorsqu'elle recouvre à la fois l'objet et le fond (cf cadre bleu).

Plusieurs mesures de distance en fonction de l'erreur de prédiction sont tracées sur la figure 4. Ces courbes montrent qu'une distance robuste (dans le cas présent, la distance L1 saturée) conduit à un profil très pointu peu approprié pour prendre en compte la $2^{\text{ième}}$ considération. De plus, la distance robuste ne satisfait pas très bien non plus le $3^{\text{ième}}$ point, puisqu'une particule



a b c

FIG. 4: *a) Profil de distance, en fonction de l'erreur (en unité de pixel) par rapport à la correspondance exacte (erreur=0). La taille des patchs considérés est 40×40, et les valeurs ont été moyennées sur 20 patchs differents. Distances : (rouge,+) L2 (vert,o) L1 saturé (bleu,×) Haussdorf. b) et c) Profil de distance pour une particule qui couvre originellement seulement l'objet (vert,×) ou en partie l'objet (60%) et le fond (rouge,o). b) distance L1 saturé c) distance L2. La différence entre les mouvements du fond et de l'objet est de 4 pixels. L'erreur est considérée comme nulle lorsque la particule a le même mouvement que l'objet.*

couvrant originellement entièrement l'objet mais avec une petite erreur de prédiction (e.g. 1 pixel) reçoit une distance similaire à une particule couvrant originellement l'objet et le fond mais dont la prédiction correspond parfaitement à la position de l'objet (cf figure 4b). Vis à vis de ces critères, la norme L2 est plus appropriée[2]. Nous avons donc choisi la distance L2, qui correspond à un bruit additif gaussien dans l'équation (6),

$$\mathrm{d}_c(\tilde{I}_{c_k}, \tilde{I}_{c_{k-1}}) = \sum_{\mathbf{r}\in R} \rho(\tilde{I}_{c_k}(\mathbf{r}) - \tilde{I}_{c_{k-1}}(\mathbf{r})) \text{ with } \rho(x) = \|x\|^2$$

où $\lambda_c = \frac{1}{2\sigma_c^2}$ et σ_c représente l'écart type du bruit.

Soulignons ici que la méthode ne fait pas du template matching, comme dans [5]. Aucun template d'objet n'est défini hors-ligne ou au début de la séquence, et le tracker ne maintient pas à jour un unique template. Ainsi, le terme de correlation n'est pas objet-spécifique (excepté au travers de la définition de la région de référence R). Une particule localisée sur le fond de l'image peut recevoir une vraisemblance importante si le mouvement prédit est en accord avec le mouvement du fond.

4 Résultats, conclusion

Nous présentons deux résultats de suivi de tête. Dans les deux cas, le tracker est initialisé à la main.

Le premier exemple, figure 5, illustre l'apport de la méthode en cas d'ambiguïtés. Malgré la présence d'une forte texture dans le fond, introduisant un bruit très important dans les mesures de contour, la tête est correctement suivie en dépit des mouvements de caméras et de la tête, de la variation d'apparence de la tête, et d'occlusions partielles. Quel que soit le nombre de particules utilisées et la variance dans le modèle dynamique, l'utilisation du seul model d'objet ne permet jamais d'effectuer un suivi correct au delà de l'instant t_{12}.

Dans la seconde séquence, figure 6, la personne suivie effectue plusieurs tours sur elle même de sorte que l'utilisation de modèles d'apparence, basés sur des histogrammes de couleur ou des templates par exemple, sont inappropriés. Dans cette séquence, le tracker basé sur la forme uniquement est perturbé vers l'instant 60 par différents facteurs (mouvement vertical abrupt de la caméra, absence de contours de la tête lorsque

2. Le besoin du point 3 est discutable : en cas d'occlusion partielle, une distance robuste pourrait être plus appropriée.

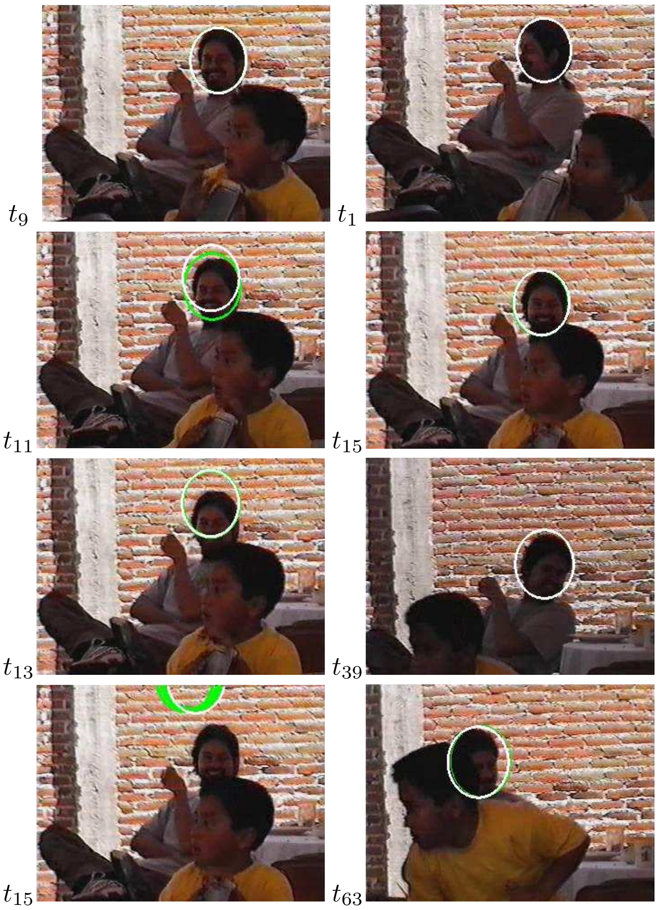


FIG. 5: *Colonne de gauche : approche utilisant le modèle de contour uniquement. Colonne de droite : notre approche avec terme de corrélation. L'ellipse blanche indique le mode principal, une verte un mode secondaire.*

FIG. 6: *Colonne de gauche : contour uniquement. A droite : approche avec terme de corrélation supplémentaire.*

celle-ci se trouve devant la bibliothèque) puis se perd. Avec notre approche, la rotation de la tête est bien suivie[3], les perturbations sont plus faibles, et le suivi s'effectue correctement sur le reste de la séquence.

Nous avons proposé une nouvelle méthode pour le suivi visuel par méthode Monte Carlo séquentiel. Celle-ci prend en compte la corrélation temporelle qui existe entre deux images successives d'un même objet par l'intermédiaire de la fonction de vraisemblance. Elle peut s'interpréter comme la prise en compte implicite de mesures de mouvement et s'avère très utile pour supprimer des ambiguïtés locales lors de l'utilisation de modèles d'objets génériques basés sur la forme ou la couleur.

Références

[1] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, pp 142–151, 2000.

[2] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[3] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *4th European Conf. Computer Vision*, volume 1, pp 343–356, 1996.

[4] Y. Raja, S. McKenna, and S. Gong. Colour model selection and adaptation in dynamic scenes. In *5th European Conference on Computer Vision*, pp 460–474, 1998.

[5] J. Sullivan and Rittscher J. Guiding random particles by deterministic search. In *ICCV*, pp 323–330, 2001.

[6] Hai Tao, Harpreet S. Sawhney, and Rakesh Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):75–89, 2001.

[7] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. 8^{th} IEEE Int. Conf. Computer Vision*, Vancouver, July 2001.

3. Ceci est un cas difficile pour la méthode, qui reçoit des informations contradictoires : l'intérieur de la tête indique un mouvement vers la droite alors que le contour extérieur de la tête reste statique.