Méthode de suivi d'objets basée sur des trajectoires temporelles de points d'intérêt

Vincent Garcia, Éric Debreuve, Michel Barlaud

Laboratoire I3S, Université de Nice-Sophia Antipolis / CNRS - UMR 6070,
Bât. Algorithmes/Euclide B, BP 121, 2000, route des Lucioles
06903 Sophia Antipolis Cedex, France
garciav@i3s.unice.fr, debreuve@i3s.unice.fr
barlaud@i3s.unice.fr

Résumé – Ce papier traite du problème de suivi de régions d'intérêt (RI) dans une vidéo, à partir d'une initialisation manuelle, pour des applications telles que la vidéo surveillance ou la post-production cinématographique. Plus exactement, nous étudions le suivi de RI réalisé à partir de trajectoires temporelles de points d'intérêt. Les méthodes classiques utilisant de telles trajectoires sont faussées dès lors qu'elles travaillent à partir d'un nombre trop faible d'observations. Nous proposons dans ce papier d'augmenter ce nombre en estimant le mouvement de la RI sur un groupe d'images. Nous montrons sur des exemples réels que notre approche est précise, fiable, et qu'elle permet d'améliorer la qualité de suivi par rapport à une estimation de mouvement sur seulement deux images.

Abstract – This paper deals with region-of-interest (ROI) tracking in a video, from an hand-edited initialization, for applications such as video surveillance or cinematographic post-production. To be more precise, we study the ROI tracking using temporal trajectories of interest points. Classical methods using these trajectories are inaccurate when applied on an insufficient number of observations. In this paper, we propose to increase the number of observations by performing the ROI motion estimation process on a group of frames. We show on real examples that the proposed method is accurate, reliable, and allows to increase the tracking quality in comparison of estimating the motion on two frames.

1 Introduction

Le problème du suivi d'objets (ou de région d'intérêt) est un thème de recherche très actif dans le domaine du traitement de vidéos. C'est un processus de bas niveau requis pour de nombreuses applications telles que la vidéo surveillance ou la post-production cinématographique. Dans ce papier, nous étudions le suivi d'objets réalisé à partir des trajectoires temporelles de points d'intérêts [1, 2]. Le suivi est initialisé sur la première image par un contour paramétrique défini manuellement.

Quelques papiers dans la littérature [3, 4] traitent du suivi d'objets basé sur l'utilisation de points d'intérêt. Ces méthodes reposent sur un même schéma : dans un premier temps, les points d'intérêt sont extraits sur les images de la vidéo ; dans un deuxième temps, les points d'intérêt sont appariés en utilisant l'information contenue dans le voisinage local de chaque point (cette information est alors appelée descripteur). Les appariements trouvés sont alors utilisés pour détecter la position de l'objet dans la nouvelle image. Leur nombre varie selon la taille et la nature de l'objet à suivre. Un nombre trop faible peut alors "empêcher" une bonne estimation du mouvement de l'objet.

Dans ce papier, nous proposons d'augmenter le nombre d'appariements considérés en utilisant un groupe d'images. Nous montrerons que cette augmentation permet d'améliorer la précision et la robustesse de l'estimation de mouvement aux données aberrantes.

Le papier que nous présentons est organisé de la manière suivante : la section 2 expose la méthode de suivi proposée. La section 3 présente et discute quelques résultats de suivi. Enfin, la section 4 conclut.

2 Estimation du mouvement de la région d'intérêt sur un groupe d'images

Dans cette section, nous proposons une méthode de suivi reposant sur l'analyse de trajectoires temporelles de points d'intérêt appelées *tracks*. Nous faisons l'hypothèse que le mouvement global d'objets suffisamment larges peut être estimé à partir des tracks. Dans un premier temps, nous expliquerons la construction des tracks avant de détailler dans un deuxième temps la méthode de suivi proposée.

2.1 Construction des *tracks*

Un point d'intérêt (PI) [1, 2], également appelé *point clé* ou *point saillant*, est usuellement défini comme étant un pixel d'une image remarquable par la présence de propriétés locales discriminantes (*e.g.*, coin : intersection de deux contours ayant des directions différentes). Combinés avec un descripteur [5, 6] (information décrivant le voisinage local de chaque PI), les PI de deux images différentes sont appariés pour former des couples de points. En choisissant

un extracteur de PI et un descripteur adaptés, les appariements de PI sont utilisés pour résoudre des problèmes difficiles [7, 8, 9]. Un *track* est la trajectoire temporelle d'un point d'intérêt construit en mettant bout à bout les appariements de PI.
Les propriétés discriminantes des descripteurs diffèrent selon l'information utilisée pour décrire le voisinage du PI. Par exemple, l'utilisation d'une fenêtre contenant l'information de niveau de gris ou de couleur est spatialement discriminante mais n'est pas robuste aux rotations. *A contrario*, la distribution de probabilité des niveaux de gris est robuste aux rotations et aux changements d'échelle (dans une moindre mesure) mais n'est pas spatialement discriminante. Nous utiliserons dans cet article l'extracteur de Harris-Stephens [1], connu pour ses excellentes propriétés d'extraction [2], combiné avec une fenêtre locale circulaire contenant les informations de niveau de gris. L'appariement étant réalisé sur des images consécutives (mouvement faible), le descripteur utilisé a d'excellentes propriétés discriminantes.
Plus précisément, les *tracks* sont construits de la manière suivante. Premièrement, les PI sont extraits indépendamment dans chaque image de la vidéo. Ensuite, le descripteur est calculé pour chaque PI. Les PI de l'image I^j sont alors appariés avec les PI de l'image I^{j+1} en mettant en correspondance les descripteurs (norme L_1 de la différence entre descripteurs) et en utilisant une validation croisée. Le mouvement entre deux images consécutives est supposé suffisamment faible pour réaliser l'appariement dans une fenêtre de recherche. Enfin, les paires de PI sont concaténées de manière à former plusieurs ensembles de PI appelés *tracks*. Les tracks ne sont généralement pas définis sur l'intégralité de la séquence vidéo mais sur un sous-ensemble d'images. Par la suite, nous utiliserons la notation suivante : un track T_k définit sur l'intervalle $[I^i, I^j]$ est une ensemble de PI $T_k = \{t_k^i, ..., t_k^j\}$. Les tracks sont la base de la méthode que nous proposons.

2.2 Estimation du mouvement de la région d'intérêt

Soient V une vidéo composée de n images $I^1, \cdots, I^n$, et C^1 le contour de la région d'intérêt (RI) initialisé manuellement sur l'image I^1. Le problème de suivi consiste à calculer les contours $C^2, \cdots, C^n$ à partir de C^1.

2.2.1 Estimation sur deux images

La méthode que nous proposons repose sur l'hypothèse suivante : le mouvement de la RI peut être déduit à partir du mouvement des PI appartenant à la RI. En d'autres termes, au temps $j+1$, la RI est guidée au cours du temps par les tracks appartenant au tube temporel formé par le contour C^1 définit manuellement et par les contours $C^2, ..., C^j$ précédemment calculés.
Pour exposer la méthode proposée, nous supposerons dans la suite que les contours $C^1, ..., C^m$ sont déjà définis. Le problème de suivi se résume alors à calculer C^{m+1} à partir de C^m et des tracks.
Premièrement, chaque track T_k appartenant au tube temporel formé par $\{C^1, ..., C^m\}$, et au moins défini sur les images I^m et I^{m+1}, est sélectionné. Deuxièmement, les paires de PI $\{t_k^m, t_k^{m+1}\}$ sont extraites des tracks sélectionnés. Troisièmement, une matrice de mouvement affine M est estimée à partir de ces paires en utilisant le M-estimateur [10] suivant :

$$M = \arg\min_M \sum_k f(\|M.t_k^m - t_k^{m+1}\|), \qquad (1)$$

où t_k^m et t_k^{m+1} sont donnés en coordonnées projectives, M est une matrice 3×3 de mouvement affine, f une fonction de coût, et $\|.\|$ la norme Euclidienne. Dans ce papier, nous utilisons f comme étant la fonction valeur absolue, et la minimisation est effectuée en utilisant la méthode du simplex [11]. Nous choisissons un M-estimateur car, contrairement à la méthode classique de minimisation des moindres carrés, l'estimation des paramètres est précise et robuste aux données aberrantes (appariements incorrects de PI). Le contour C^j, défini sur l'image I^j, est un ensemble de points d'échantillonnage $\{p_1^j, \cdots, p_l^j\}$. Ainsi, le contour C^{m+1} est déduit en appliquant M aux points d'échantillonnage de C^m :

$$p_i^{m+1} = M.p_i^m, \ \forall i \in [1, l], \qquad (2)$$

où p_i^m et p_i^{m+1} sont donnés en coordonnées projectives.

2.2.2 Estimation sur un groupe d'images

Les M-estimateurs, comme la plupart des méthodes statistiques, nécessitent un nombre important de données pour fournir une estimation précise et robuste des paramètres (paramètre de la matrice de mouvement M dans notre cas). Le suivi d'objets de petite taille peut faire intervenir un nombre trop faible de tracks pour garantir une estimation précise du mouvement (souvent moins de 10 tracks sélectionnés d'où moins de 10 paires de PI utilisées pour l'estimation des paramètres).
Pour remédier à cette pénurie d'information, nous pouvons changer la sensibilité de l'extracteur de PI ce qui augmente le nombre de PI extraits et le nombre de tracks. Cependant, ceci diminuerait la qualité des points extraits et, par conséquent, diminuerait la qualité des tracks.
Nous proposons de considérer plus de paires de PI (sans changer le nombre de tracks) en les extrayant non plus sur les images I^m et I^{m+1}, mais sur un ensemble d'images centré sur les images $\{I^m, I^{m+1}\}$. Nous faisons l'hypothèse suivante : dans une séquence vidéo, le mouvement des points (PI et points d'échantillonnage) est constant sur un ensemble d'images (ou inversement, la taille de l'ensemble d'images peut être choisi tel que l'hypothèse soit raisonnable). Nous supposerons donc que le mouvement des points est constant sur un ensemble de G images (G étant un nombre pair) centré sur les images $\{I^m, I^{m+1}\}$. Ainsi, au plus $G-1$ paires de PI sont extraites pour chaque track sélectionné :

$$\{t_k^{m-g}, t_k^{m-g+1}\}, \{t_k^{m-g+1}, t_k^{m-g+2}\}, \cdots$$
$$\cdots, \{t_k^m, t_k^{m+1}\}, \cdots, \{t_k^j, t_k^{j+1}\}, \cdots, \{t_k^{m+g}, t_k^{m+g+1}\}$$

où g est la demi-taille de l'ensemble d'images : $g = \frac{G}{2}$. En extrayant les paires de PI sur l'intervalle $[I^{m-g}, I^{m+g+1}]$,

nous supposons implicitement que les tracks définis dans les contours $C^1, \cdots, C^m$ seront définis dans les contours $C^{m+1}, \cdots, C^{m+g+1}$. Nous proposons de pondérer les paires de PI en donnant plus d'importance aux paires temporellement proche des images I^m et I^{m+1}. La matrice de mouvement M est alors estimée en minimisant le M-estimateur [10] pondéré suivant :

$$M = \arg\min_M \sum_k \sum_{j=m-g}^{m+g} \delta_j . f(\|M.t_k^j - t_k^{j+1}\|), \quad (3)$$

où t_k^j et t_k^{j+1} sont donnés en coordonnées projectives, M est une matrice 3×3 de mouvement affine, $f(x) = |x|$, $\|.\|$ la norme Euclidienne, et la pondération temporelle δ_j pour une paire de PI $\{t_k^j, t_k^{j+1}\}$ est donnée par :

$$\delta_j = \psi(|m - j|), \quad (4)$$

où ψ est une fonction positive, monotone décroissante et définie sur $\mathbb{R}^+$. Nous choisissons ψ comme étant une fonction Gaussienne. Cependant, toute fonction dérivée des fonctions de régularisation classiques [12] peut être utilisée à la place. Les pondérations temporelles δ_j sont différentes pour chaque paire extraites du track T_k mais similaires pour chaque k. Finalement, le contour C^{m+1} est déduit en appliquant M aux points d'échantillonnage de C^m comme dans l'équation (2).

3 Expérimentations

Cette section se compose de deux parties distinctes. Dans la première, nous montrerons l'intérêt de l'utilisation d'un groupe d'images pour l'estimation de la matrice de mouvement. Ceci représente la contribution majeure de ce papier. Dans un second temps, nous confronterons la méthode proposée à deux méthodes de suivi classiques.

3.1 Amélioration apportée par l'utilisation d'un groupe d'images

Nous montrons ici que l'augmentation du nombre de paires de PI permet de diminuer l'erreur de suivi. Cette erreur, exprimée en pourcentage de pixels mal classés, est obtenue en estimant la différence symétrique normalisée entre le masque du contour calculé et le masque du contour exact défini manuellement. L'étude porte sur deux séquences vidéo réelles.

La première est la séquence *Crew* de 80 images au format CIF (CIF=288 $\times$ 352 pixels) (voir figure 1). Le minimum global de la courbe présentée sur la figure 1 nous indique que l'estimation du mouvement est optimale pour un groupe contenant 8 images. Ce minimum représente le compromis optimal entre le nombre de paires de PI extraites et la validité de l'hypothèse de constance du mouvement : avant ce minimum, le nombre de paires de PI n'est pas suffisant pour estimer précisément les paramètres de la matrice de mouvement ; après ce minimum, l'hypothèse n'est plus vérifiée. Nous remarquons également sur la figure 1 que l'utilisation d'un groupe d'images (ici 8 images) permet de diviser par 2 l'erreur de suivi par rapport à une approche simple (utilisation de 2 images).

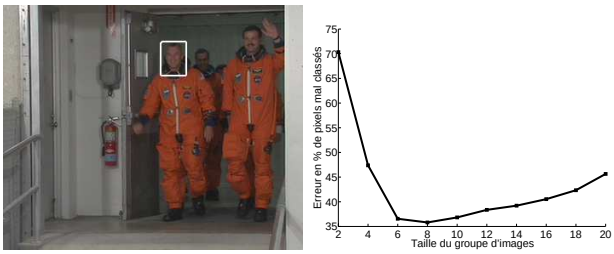

FIG. 1 – Évolution du pourcentage de pixels mal classés sur la séquence *Crew* de 80 images au format CIF.

La seconde séquence vidéo est tirée du film « Le Voyage en Arménie »[1] (voir figure 2). Cette séquence HD (HD=1920 $\times$ 1080 pixels) de 40 images permet de tester notre méthode sur des données cinématographiques réelles. Le contour utilisé ici est le contour approximatif de la voiture. Le mouvement de la voiture est quasiment une pure translation (variance de la norme des vecteurs mouvement = 0.15 pixels pour un mouvement de 2 pixels, et variance de l'orientation = 3 degrés). La courbe présentée sur la figure 2 indique que la taille optimale du groupe d'image est de 30. Cette taille élevée confirme la remarque précédente sur la nature du mouvement. L'utilisation d'un groupe d'images (ici 30 images) permet de diviser par 2.5 l'erreur de suivi par rapport à une approche simple (utilisation de 2 images).

En résumé, l'utilisation d'un groupe d'images pour l'estimation du mouvement permet d'améliorer la qualité du suivi par rapport à l'estimation sur deux images. Le choix de la taille du groupe d'images utilisée dépend naturellement du mouvement de l'objet dans la séquence vidéo.

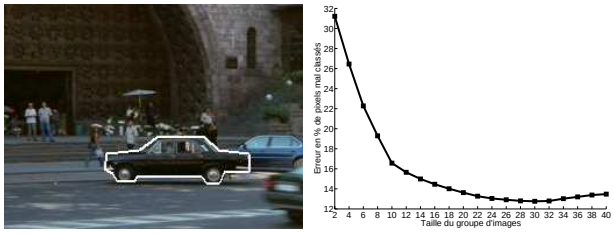

FIG. 2 – Évolution du pourcentage de pixels mal classés sur la séquence *Arménie* de 40 images au format HD.

3.2 Comparaison de la méthode proposée à deux méthodes de suivi classiques

Dans cette section, nous comparons la méthode proposée à deux méthodes classiques. La première est une simple méthode de bloc-matching [13] utilisant une recherche sous-optimale [14]. Pour cette méthode, le contour C^1, édité manuellement et correspondant à la RI sur la première image, est utilisé comme bloc initial. Le suivi est réalisé en détectant sur les images $[I^2, I^n]$ les blocs qui minimisent un critère de similarité (somme de la valeur absolue de la différence des luminances). La seconde

[1] Mentions légales : film "Le Voyage en Arménie" - Réalisation : Robert GUEDIGUIAN - Production : AGAT FILMS & Cie

FIG. 3 – Comparaison entre l'estimation du mouvement à partir des tracks sur deux images (image de gauche) et l'estimation du mouvement sur un groupe de 30 images (image de droite). L'utilisation d'un groupe d'image améliore la qualité du suivi de l'objet.

méthode utilisée à titre de comparaison est une méthode reposant sur le *mean-shift* [15, 16]. Pour cette expérimentation, les trois méthodes sont comparées sur la séquence CIF de 20 images intitulée *Football*. Le contour initial est un rectangle englobant le casque d'un footballeur américain et est strictement le même pour chacune des trois méthodes (voir figure 4). Sur cette séquence, seule la méthode proposée permet de suivre précisément la RI, les deux autres méthodes étant perturbées par la présence de pixels du fond dans le contour initial. La méthode proposée apparaît donc comme étant une méthode de suivi précise, très fiable, et robuste aux données aberrantes.

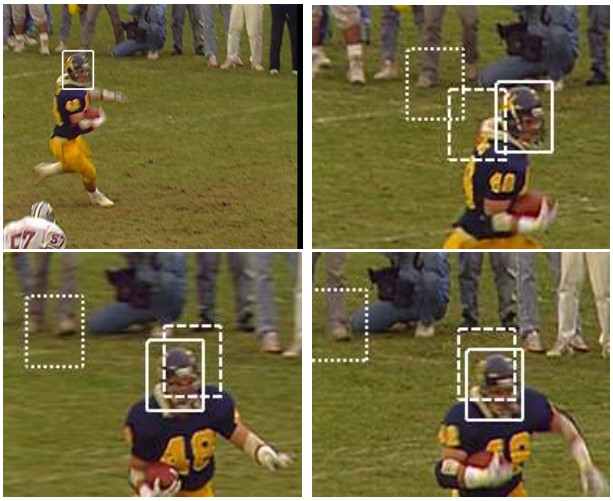

FIG. 4 – Comparaison des trois méthodes de suivi sur la séquence *Football* : bloc-matching (ligne en pointillés), mean-shift (tirets), et la méthode proposée (ligne pleine). Les images de la séquence exposées sont I^1, I^7, I^{14}, et I^{20} (de gauche à droite et de haut en bas).

4 Conclusion

Nous avons proposé une méthode de suivi reposant sur l'estimation du mouvement de régions d'intérêt à partir de trajectoires temporelles. La méthode permet d'utiliser toute forme de contour de manière à bien représenter la forme de la région d'intérêt. La contribution majeure de ce papier repose sur l'utilisation d'un ensemble d'images permettant d'augmenter le nombre d'observations ce qui améliore la qualité du suivi et la robustesse aux données aberrantes. La méthode proposée fournit un suivi précis sur des séquences réelles.

Références

[1] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of The Fourth Alvey Vision Conference*, Manchester, UK, 1988, pp. 147–151.

[2] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. Journal on Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[3] R. Trichet and B. Mérialdo, "Probabilistic matching algorithm for keypoint based object tracking using a Delaunay triangulation," in *Int. Workshop on Image Analysis for Multimedia Interative Services*, Santorini,Greece, June 2007.

[4] J. Hoey, "Tracking using flocks of features, with application to assisted handwashing," in *Proc. of the British Machine Vision Conference*, Edinburgh, Scotland, August 2006, vol. 1, pp. 367–376.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] F. Tonnin, P. Gros, and C. Guillemot, "Analysis of multiresolution representations for compression and local description of images," in *Int. Conf. on Visual Information and Information Systems*, July 2005, pp. 234–246.

[7] S. Bres and JM. Jolion, "Detection of interest points for image indexation," in *Int. Conf. Visual Information and Information Systems*, 1999, pp. 427–434.

[8] Ch. Wolf, J. M. Jolion, W. Kropatsch, and H. Bischof, "Content based image retrieval using interest points and texture features," in *IEEE Int. Conf. on Pattern Recognition*, 2000.

[9] D. G. Lowe, "Robust model-based motion tracking through the integration of search and estimation," *Int. Journal of Computer Vision*, 1992.

[10] C. Gourieroux and A. Monfort, *Statistics and Econometric Models*, vol. 1, Cambridge University Press, 1995.

[11] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder-mead simplex algorithm in low dimensions," *SIAM Journal on Optimization*, vol. 9, pp. 112–147, 1998.

[12] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. in Image Processing*, vol. 6, no. 2, pp. 298–311, February 1997.

[13] T. Koga, K. Linuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion compensated interframe coding for video conferencing," *Proc. Nat. Telecommun. Conf.*, 1981.

[14] S. Zhu and K.K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. On Image Processing*, vol. 9, no. 2, February 2000.

[15] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, 2000, pp. 142–151.

[16] Robert Collins, Xuhui Zhou, and Seng Keat Teh, "An open source tracking testbed and evaluation web site," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005), January 2005*, January 2005.