Extraction d'objets-clés pour l'analyse de vidéos

Jérémy Huart, Pascal Bertolino

GIPSA-Lab, INPG-CNRS
ENSIEG, Domaine universitaire, Grenoble, France
jeremy.huart@gipsa-lab.inpg.fr, pascal.bertolino@gipsa-lab.inpg.fr

Résumé – Cet article propose une méthode de représentation de plans vidéo provenant d'une caméra mobile. L'approche est fondée sur les objets contenus dans la vidéo. Elle utilise une extraction des régions du premier plan capables de représenter des objets d'intérêt sémantiques. Cependant, les régions du premier plan extraites en chaque image par compensation de mouvement ne sont pas toujours représentatives de l'entité dont elles proviennent. Un filtrage et une classification de ces régions nous permet de retenir uniquement la plus représentative de chaque objet réel. C'est ce que nous appelons les *objets-clés*.

Abstract – This paper discusses object-based representation of video shots acquired by mobile camera. Our approach uses an extraction of foreground regions capable of representing semantic objects of interest. However, foreground regions, extracted by motion compensation, are not always representative of the entity they stem from. A filtering and a clustering of these regions allow us to retain only the most representative of each real object in the shot. These representations of quality are called key-objects.

1 Introduction

La description compacte du contenu d'une vidéo est actuellement une tâche rendue difficile par la très grande quantité de données qu'elle contient. Une représentation classique d'un plan peut être réalisée par une sélection appropriée d'une ou plusieurs images-clés à l'aide de critères tels que la couleur, le mouvement, ... Une synthèse des principales techniques pour l'extraction des images-clés est disponible dans [1]. Récemment, quelques travaux s'intéressent à des représentations fondées sur les objets [2, 3, 4, 5]. Deux familles d'approches peuvent être distinguées : tout d'abord celles qui sélectionnent dans le plan les images dans lesquelles la recherche d'objets sera effectuée [2, 6]. La deuxième approche consiste à extraire tout au long du plan des régions clés afin de collecter de l'information sur ces régions et en déduire une représentation du plan [7, 8, 4]. C'est dans cette seconde approche que notre méthode se positionne.

Dans notre approche, les régions sont tout d'abord grossièrement extraites par compensation du mouvement dominant. Ensuite une segmentation [9] réalisée uniquement en périphérie de ces régions permet d'obtenir des masques raffinés, correspondant bien aux contours véritables de l'objet réel appelé *objet d'intérêt* par la suite (OI). Notons qu'une partie seulement de l'OI peut avoir un mouvement apparent. Il peut également être partiellement ou temporellement occulté. Ainsi, il est souvent impossible d'extraire dans chaque image un objet vidéo (VOP) totalement représentatif de l'OI. Les régions extraites sont donc souvent des sous-objets vidéos (S-VOPs) pas nécessairement représentatifs de l'OI (*cf.* fig. 1).

A partir d'une vidéo préalablement découpée en plans, notre méthode extrait un ensemble d'occurrences (ou S-VOPs) pour chaque objet d'intérêt (*cf.* fig. 2). L'occurrence la plus représentative est appelée *objet-clé*. La chaîne

Fig. 1 – Exemple de S-VOPs extraits d'un OI non rigide (plan-séquence *children*) à l'aide de notre méthode

d'extraction d'un objet-clé se décompose en plusieurs étapes :

1. Extraction des S-VOPs
2. Rejet des S-VOPs de mauvaise qualité
3. Classification couleur des S-VOPs : une classe par S-VOP (S-VOP générateur)
4. Suppression dans chaque classe des S-VOPs dont la trajectoire est non cohérente avec le S-VOP générateur
5. Fusion des classes similaires pour obtenir une classe par OI
6. Rejet des classes temporellement non significatives
7. Sélection de l'objet-clé pour chaque classe

Chaque étape de ce procédé peut être perçue comme une boîte noire fournie avec un nombre fini d'entrées/sorties. Ainsi, chaque boîte peut éventuellement être remplacée par une autre plus efficace ou dédiée à une application donnée. La section suivante détaille ces différentes étapes.

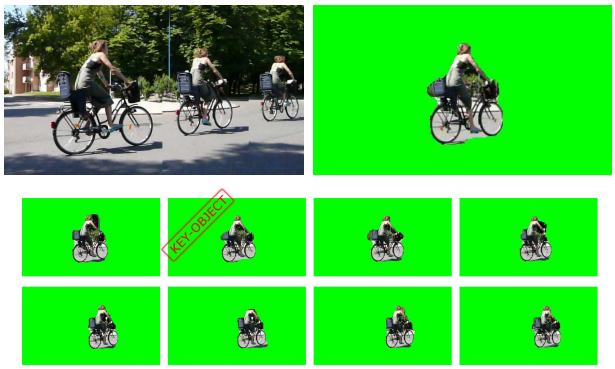


FIG. 2 – Exemple de traitement utilisant notre méthode. En haut à gauche : montage de 3 images pour donner un aperçu du plan. En haut à droite : l'objet-clé extrait. En bas : quelques S-VOPs extraits

2 La chaîne d'extraction des objets-clés

2.1 Extraction de l'ensemble des S-VOPs

L'extraction des S-VOPs peut être obtenue avec toute technique qui fournit pour chaque image un ensemble de masques d'entités en mouvement. Nous utilisons une technique rapide qui calcule un modèle de mouvement global paramétrique par image [10], couplée avec une segmentation raffinée du contour des objets [9]. Cette étape fournit pour chaque image du plan un ensemble (éventuellement vide) de masques (S-VOPs) dont les contours ont pour but d'être fidèle à la forme des objets. Notons qu'ici, aucun suivi d'objet n'est réalisé.

2.2 Rejet des S-VOPs de mauvaise qualité

Les S-VOPs non représentatifs d'un OI sont supprimés de l'ensemble des S-VOPs. Le rejet est fondé sur 2 critères (détaillés dans les paragraphes suivants) : le premier fait l'hypothèse qu'un S-VOP de qualité a une forme compacte, alors que le second requiert une bonne correspondance entre le bord du S-VOP et les contours de l'OI.

2.2.1 Compacité : une caractéristique discriminante

L'extraction d'objet en mouvement est souvent sujet aux « fuites » de l'objet vers le fond. Les régions résultantes sont alors fines et/ou allongées et donc peu compactes. La compacité C d'un S-VOP s est donnée par le facteur de forme :

$$C(s) = \frac{\text{Perimetre}(s)^2}{4\Pi \times Surface(s)} \quad (1)$$

$C \in [1,\ \infty]$. Dans la majorité des plans traités, nous avons observé un mode principal pour une valeur de C proche de 1 représentant des régions compactes. Un seuil empirique fixé à 2.5 permet d'exclure les S-VOPs peu compacts.

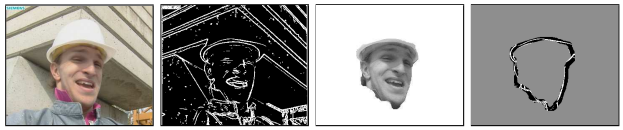
2.2.2 Évaluation de la qualité des contours

La qualité Q d'un S-VOP s est donnée par le degré de recouvrement entre le contour de son masque et les contours de l'OI : Soit z le contour épais de s et e les contours obtenus par un seuillage adaptatif du gradient de Sobel dans l'image originale[1] :

$$z(s) = Dilatation_\epsilon(s) \setminus Erosion_\epsilon(s) \quad (2)$$

ϵ est un élément structurant de rayon égal à quelques pixels (typiquement 6). s est rejeté lorsque $Q(s) < T$. T est obtenu adaptativement à l'aide d'une modélisation par une Gaussienne de la distribution des $Q(s)$.

$$Q(s) = \frac{card(e \in z(s))}{Surface(z(s))} \quad (3)$$



(a) Image originale (b) Gradient seuillé (c) S-VOP (d) e (en blanc) et z (en noir)

FIG. 3 – Mesure de qualité d'un S-VOP par recouvrement du masque et des contours

2.3 Classification couleur des S-VOPs

Le but est de répartir les n S-VOPs extraits en m classes (en général, $n \gg m$) représentant les m OI. Comme m est *a priori* inconnu, une classification en 2 étapes sur la couleur est utilisée : n classes couleur sont construites, chaque classe comprenant initialement un S-VOP (appelé S-VOP générateur). Pour chaque classe, tous les S-VOPs similaires en couleur au S-VOP générateur sont rajoutés à la classe. La similarité est calculée sur le recouvrement de mélanges de Gaussiennes modélisant la distribution couleur des S-VOPs.

Afin de quantifier ce recouvrement nous utilisons un critère présenté dans [11] : 2 Gaussiennes $\mathcal{N}(\mu_1, \Sigma_1)$ et $\mathcal{N}(\mu_2, \Sigma_2)$ sont ***c**-séparées* si :

$$\|\mu_1 - \mu_2\| \geqslant \mathbf{c} \cdot \sqrt{2 \cdot max(\lambda_{max}(\Sigma_1), \lambda_{max}(\Sigma_2))} \quad (4)$$

Avec $\lambda_{max}(\Sigma_1)$ et $\lambda_{max}(\Sigma_2)$ les valeurs singulières les plus élevées des matrices de covariance Σ_1 et Σ_2.

2 Gaussiennes 1-*séparées* ou $^1/_2$-*séparées* se recouvrent significativement. Ce recouvrement permet de comparer les mélanges de Gaussiennes. Soient m_i et m_j, 2 mélanges de Gaussiennes modélisant s_i et s_j. m_i est inclus dans m_j si et seulement si chaque Gaussienne de m_i est au moins 1-séparée d'une des Gaussiennes de m_j. Si une inclusion est vérifiée entre m_i et m_j alors s_i et s_j sont regroupés dans la même classe couleur. L'utilisation de l'inclusion de mélanges permet de regrouper dans une même classe

[1] $A \setminus B = \{x \in A \text{ et } x \notin B\}$.

les S-VOPs provenant de différentes sous-parties d'un OI ayant des couleurs communes.

2.4 Contrôle de trajectoire

Afin de prendre en compte l'information spatio-temporelle au sein de chaque classe, les S-VOPs dont la trajectoire n'est pas compatible avec celle construite à partir du S-VOP générateur sont supprimés. Cette contrainte permet de différencier facilement des objets similaires en couleur ayant des trajectoires croisées ou des objets qui ont des trajectoires identiques mais à des moments différents.

La trajectoire est calculée en utilisant la position et le mouvement compensé du centre de gravité G_{ref} du S-VOP de référence s_{ref}. Le contrôle consiste à rechercher itérativement le S-VOP correspondant dans les images voisines de la référence. La toute première référence est le S-VOP s_{gen}. La recherche est effectuée dans chaque image du plan en 2 étapes : de s_{gen} à la fin du plan, ensuite de s_{gen} au début du plan. La recherche est faite dans une fenêtre circulaire centrée sur la projection $G_{ref} = (x, y)$ ayant la vitesse $\vec{V} = (dx, dy)$:

$$proj(G_{ref}) = (x + dx, y + dy) \tag{5}$$

Un candidat $s_i(t)$ est temporellement cohérent avec la trajectoire de s_{ref} si et seulement si :

$$\|G_{S_{ref}} - G_{S_i(t)}\| \leqslant r \text{ et } \left\langle \vec{V}_{s_{ref}} \cdot \vec{V}_{s_i(t)} \right\rangle \geq 0 \tag{6}$$

Avec r le rayon de la fenêtre de recherche relative au rayon de s_{gen}. Dans l'équation 6, la première condition assure la cohérence spatiale et la seconde assure la conformité entre les directions du mouvement des deux S-VOPs. Dans une image donnée, la recherche est effectuée selon les règles suivantes :

1. si aucun S-VOP n'a son centre de gravité inclus dans la fenêtre de recherche, la recherche continue dans l'image suivante. La position de la fenêtre est mise à jour avec le vecteur mouvement de s_{ref}.
2. si seulement un S-VOP vérifie les conditions de l'équation 6, celui-ci appartient définitivement à la classe et devient la nouvelle référence s_{ref}.
3. si plusieurs S-VOPs vérifient les conditions de l'équation 6, ils sont tous conservés dans la classe. Le centre de gravité de l'ensemble devient le nouveau G_{ref} et $\vec{V}_{s_{ref}}$ correspond à la moyenne des vitesses.
4. chaque S-VOP dont le centre de gravité est extérieur à la fenêtre de recherche est exclu de la classe.

2.5 Fusion hiérarchique des classes

Les n classes sont ici considérées comme des ensembles. Des classes incluses entre elles ou dont l'intersection en terme de nombre d'éléments est importante doivent être fusionnées : la fusion commence par la construction d'un dendrograme (i.e. une classification hiérarchique) dans lequel les classes fusionnent itérativement deux à deux pour n'en donner finalement plus qu'une (fig. 4). A chaque itération, seule la fusion de moindre coût (concernant les deux classes les plus similaires) est réalisée.

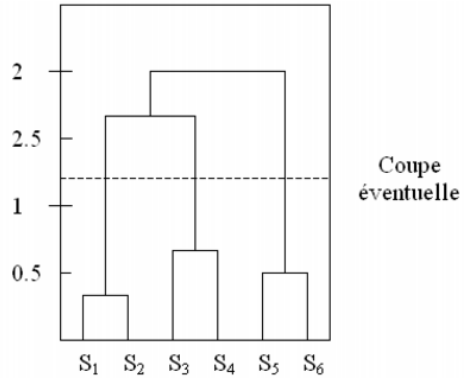


FIG. 4 – Exemple de dendrogramme construit à partir de 6 classes couleur

Le coût d'une fusion peut être évaluée selon l'intersection entre deux classes du point de vue de la théorie ensembliste. Soit δ la dissimilarité entre 2 classes c_1 et c_2 vérifiant $|c_2| \leq |c_1|$:

$$\delta = \frac{|c_2 \setminus c_1 \cap c_2|}{|c_2|} \tag{7}$$

$\delta = 1$ lorsque l'intersection est vide. $\delta = 0$ lorsque $c_2 \subset c_1$. Soit $c_3 = c_1 \cup c_2$. La dissimilarité entre la classe c_3 et la classe c est donnée par l'expression suivante :

$$\delta(c_3, c) = min[\delta(c_1, c), \delta(c_2, c)] \tag{8}$$

Le découpage final en classes (idéalement une classe par OI) est obtenu en scindant le dendrogramme selon une classification des dissimilarités. L'objectif est de maximiser l'inertie entre les deux ensembles E_i et F_i : avec E_i l'ensemble des dissimilarités faibles (≥ 0 et $< i$) et F_i l'ensemble des dissimilarités fortes ($\geq i$ et < 1, les dissimilarités supérieures à 1 sont exclues puisqu'elles traduisent des classes totalement disjointes). Soit $D = E_i \cup F_i$. m_D, m_{E_i}, m_{F_i} sont les moyennes de D, E_i, F_i. L'inertie est alors donnée par :

$$I_i = w_e d(m_{E_i}, m_D)^2 + w_f d(m_{F_i}, m_D)^2 \tag{9}$$

Où $w_e = |E_i|, w_f = |F_i|$ et d est la distance Euclidienne. La meilleure partition est obtenue avec la valeur i qui maximise l'inertie :

$$T = \underset{i}{\text{argmax}}(I_i) \tag{10}$$

2.6 Suppression de classes non significatives

Enfin, nous proposons d'exclure les classes qui ne sont pas temporellement significatives : une classe contient des S-VOPs dont la première et la dernière apparition sont dans les images notées i_{debut} et i_{fin}. Ceci induit la durée (en nombre d'images) et la persistance (taux d'apparition) de la classe. On émet l'hypothèse selon laquelle un OI est présent dans un plan pendant une durée significative $\Delta = i_{fin} - i_{debut}$ et est extrait $p\%$ du temps. Δ et p sont fixés empiriquement et permettent de valider ou d'exclure les classes. Il a été choisi $\Delta = 50$ *i.e.* 2 secondes et $p = 20\%$.

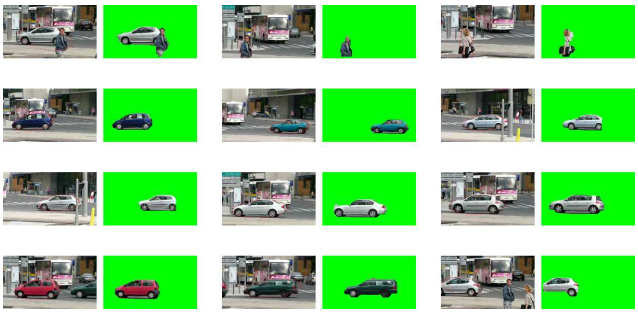


FIG. 5 – Les 12 objets-clés extraits de la vidéo *Chavant*. A gauche : les images originales. A droite : les objets-clés

(a) Résumé qui montre les images importantes du plan. En vert, l'objet-clé, en bleu les vues-clés, en rouge la zone d'occultation

(b) Quelques échantillons du suivi obtenu

FIG. 6 – Suivi contrôlé par un objet-clé et des vues-clés

2.7 Selection des objets-clés

A ce stade, chaque classe est supposée contenir plusieurs S-VOPs relatifs à un OI. Le S-VOP ayant la meilleure qualité de sa classe c (cf. l'équation 3) est considéré comme l'objet-clé. Plus précisément, l'objet-clé est le S-VOP qui maximise cette qualité dans un sous-ensemble $\widehat{c}$ de c. $\widehat{c}$ est obtenu de la façon suivante : comme Q (*cf.* équation 3) est un pourcentage, les petits S-VOPs sont privilégiés. Pour éviter ce biais, on estime l'intervalle le plus représentatif des surfaces de c : c est divisée à l'aide de l'algorithme des k-moyennes en 3 sous-ensembles disjoints correspondant aux surfaces des S-VOPs : petites, moyennes et grandes. $\widehat{c}$ est le sous-ensemble qui donne la qualité moyenne $\overline{Q}$ la plus élevée.

3 Résultats

La méthode présentée permet une extraction quasiment exhaustive et de bonne qualité des objets-clés. La vidéo *Chavant* (fig. 5) montre un carrefour filmé avec une caméra tenue à la main qui effectue un panoramique et un (de-)zoom. Le plan dure 18 secondes (soit 540 images de taille 424×240). Visuellement, on y compte clairement 14 OI en mouvement. 12 objets-clés correspondant à 12 OI ont été extraits. Parmi eux, 6 voitures qui sont de couleur gris métallisé et 2 piétons. Les 2 OI qui n'ont pas été détectés sont 2 voitures blanches qui ne sont apparues que pendant trop peu d'images et qui ont donc généré des classes temporellement non significatives. Sur un processeur Intel P4 à 2.8Ghz, l'extraction des S-VOPs (première étape) prend 20mn alors que les autres étapes prennent 10 sec. Pour conclure, notons que la méthode n'est pas sensible aux occultations ni aux changement d'échelle des objets d'intérêt. Ces objets-clés peuvent être utilisés dans de nombreuses applications qui nécessitent de connaître une ou plusieurs références de chaque objet, comme c'est le cas dans le suivi d'objets (fig. 6).

Références

[1] Y. Li, T. Zhang, and D. Tretter, "An overview of video abstraction techniques," *HP*, July 31st 2001.

[2] J.-H Oh, J. Lee, and E. Vemuri, "An efficient technique for segmentation of key object(s) from video shots," in *ITCC '03 : Proceedings of the International Conference on Information Technology : Computers and Communications*, Washington, DC, USA, 2003, p. 384, IEEE Computer Society.

[3] A. Ekin, A. Murat Tekalp, and R. Mehrotra, "Object-based video description : From low level features to semantics," in *SPIE conf. on Storage and Retrieval for Media Databases*, San Jose, CA, pp. 362-372, Jan. 2001, pp. 362–372.

[4] C. Kim and J. Hwang, "Object-based video abstraction for video surveillance systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12 (12), December 2002.

[5] H. Xu, A. A. Younis, and M. R. Kabuka, "Automatic moving object extraction for content-based applications.," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 14, no. 6, pp. 796–812, 2004.

[6] X. Song and G. Fan, "Key-frame extraction for object-based video segmentation," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2005)*, Philadelphia, PA, March 2005.

[7] J. Ruiz Hidalgo and P. Salembier, "Robust segmentation and representation of foreground key-regions in video sequences," in *Proc. IEEE Internat. Conf. on Acoustic, Speech Signal Process. (ICASSP'01)*, 2001.

[8] J. Calic and B. Thomas, "Spatial analysis in key-frame extraction using video segmentation," in *Workshop on Image Analysis for Multimedia Interactive Services*, April 2004.

[9] J. Huart, G. Foret, and P. Bertolino, "Moving object extraction with a localized pyramid," in *International Conference on Pattern Recognition*, Cambridge, UK, august 2004.

[10] S. Liu, Z. Yan, J. Kim, and C.-C. Jay Kuo, "Global/local motion-compensated frame interpolation for low-bit-rate video," *Proceedings of SPIE*, vol. 3974, pp. 223–234, april 2000.

[11] S. Dasgupta, "Learning mixtures of gaussians," in *FOCS '99 : Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, Washington, DC, USA, 1999, p. 634, IEEE Computer Society.