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

Jérémy Huart, Pascal Bertolino

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

Jérémy HUART[1], Pascal BERTOLINO

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

Thèmes choisis : 3.8 Analyse de séquences, 4.2 Segmentation et détection de ruptures
Le problème traité :
L'extraction automatique et de qualité d'objets d'intérêt dans un plan vidéo.
L'originalité :
La méthode est totalement générique et automatique. Elle sélectionne pour chaque objet en mouvement apparent l'occurrence la plus représentative qui est ensuite utilisée comme une référence pour de nombreuses applications : initialiser et contrôler un suivi, indexer par objet, construire un résumé de vidéo, ...
Les résultats nouveaux :
Les objets extraits sont quasi exhaustifs et bien représentatifs d'un plan. Pour les applications de suivi précis d'objets en mouvement, la méthode proposée permet une gestion des occultations et des changements d'apparence.

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]. Dans notre approche, les régions sont tout d'abord grossièrement extraites par compensation du mouvement dominant. Ensuite une segmentation [6] 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.

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. 1). L'occurrence la plus représentative est appelée *objet-clé*. La chaîne 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. Sélection de l'objet-clé pour chaque classe

La section suivante détaille ces différentes étapes.

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 [7], couplée avec une segmentation raffinée du contour des objets [6]. Cette étape fournit pour chaque image du plan un ensemble (éventuellement vide) de masques (S-VOPs) dont les contours ont pour but de bien épouser la forme des objets. Notons qu'ici, aucun suivi d'objet n'est réalisé.

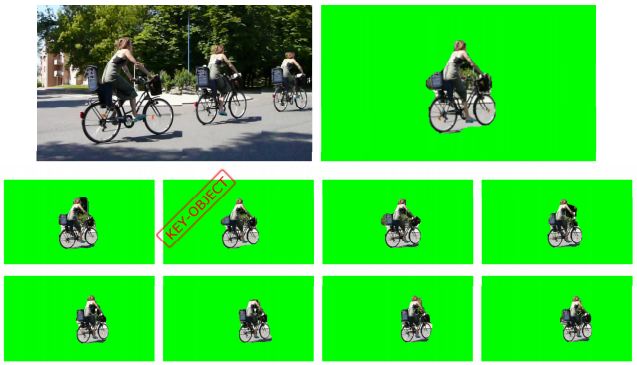


FIG. 1 – Exemple de traitement utilisant notre méthode. En haut à gauche : montage de 3 images pour donner un aperçu du plan. En haut à droite : l'objet-clé extrait. En bas : quelques S-VOPs extraits

(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

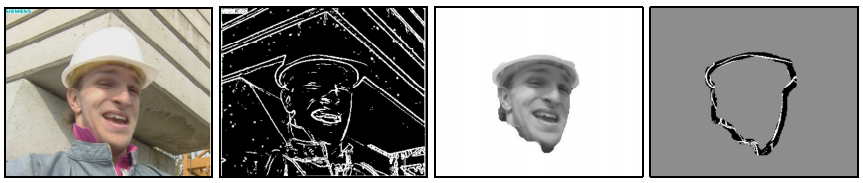
(b) Quelques échantillons du suivi obtenu

FIG. 2 – Suivi contrôlé par un objet-clé et des vues-clés

2.2 Rejet des S-VOPs de mauvaise qualité

La qualité Q_s d'un S-VOP s est donnée par le degré de correspondance entre 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] : $z(s) = Dilat_\epsilon(s) \setminus Erod_\epsilon(s)$. 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.

$$Q(s) = \frac{Card(e \in z(s))}{Area(z(s))} \tag{1}$$



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

FIG. 3 – Mesure de qualité d'un S-VOP

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 [8].

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.

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 d'éléments est importante doivent être fusionnées : la fusion commence par la construction d'un dendrogramme (i.e. une

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

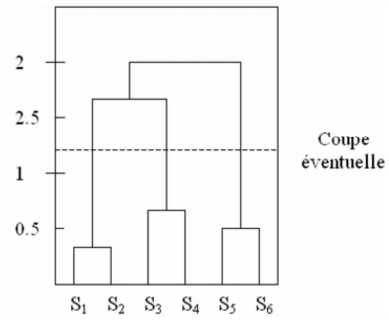


FIG. 4 – Exemple de dendrogramme construit à partir de 6 classes couleur

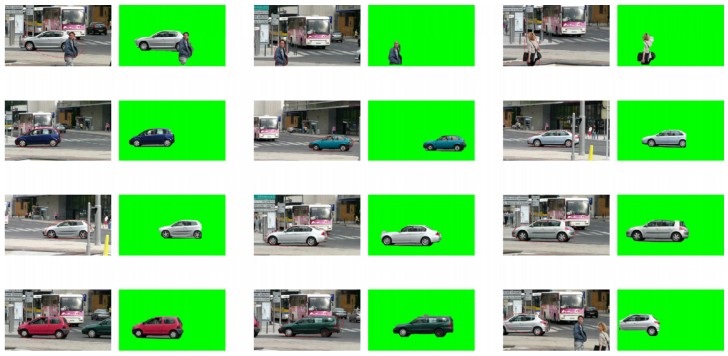


FIG. 5 – Les 12 objets-clés extraits de la vidéo *Chavant*. A gauche : les images originales. A droite : les objets-clés

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. Le découpage final en classes (idéalement une classe par OI) est obtenu en seuillant le dendrogramme avec le coût qui maximise l'inertie entre les fortes similarités et les faibles similarités du dendrogramme.

2.6 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. equ. l'équation 1) est considéré comme l'objet-clé. Plue 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 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 quasi-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. 2).

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] J. Huart, G. Foret, and P. Bertolino, "Moving object extraction with a localized pyramid," Cambridge, UK, august 2004.

[7] 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.

[8] 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.