Indexation de séquences vidéo

Indices liés au temps

Sébastien Lefèvre* — Florence Sèdes**

** LSIIT*
Parc de l'innovation, bd Sébastien Brant, BP 10413
F-67412 Illkirch
lefevre@dpt-info.u-strasbg.fr

*** IRIT*
Univ. Paul Sabatier, 118 route de Narbonne
F-31062 Toulouse
sedes@irit.fr

RÉSUMÉ. *L'objectif de cette contribition est de recenser les différents indices qui peuvent être extraits des séquences vidéo et décrire les principales méthodes permettant de fournir ces indices.*

ABSTRACT. *The aim of this paper is to inventory the features that can be extracted from video sequences and to describe the main methods enabling to identify these features.*

MOTS-CLÉS : *séquences vidéo, métadonnées, méthodes d'extraction.*

KEYWORDS: *video, segmentation, metadata, media mining methods.*

DN – 8/2004. Temps et documents, pages 41 à 48

1. Introduction : méthodes d'extraction

Les séquences vidéo sont par nature riches d'informations qu'il est intéressant d'exploiter. Pour extraire les indices les plus pertinents, des méthodes d'analyse et de traitement de séquences d'image (extension des opérations de traitement d'images) doivent être employées (Antani *et al.*, 2002 ; Correia *et al.*, 1998). Dans ce qui suit, nous allons recenser les différents indices qui peuvent être extraits des séquences vidéo et décrire les principales méthodes permettant de fournir ces indices. Nous avons choisi de présenter ces indices et les méthodes qui leur sont associées en deux parties, selon qu'ils soient relatifs à des caractéristiques globales de la scène ou aux objets.

2. Indices relatifs à la scène

Nous détaillons ici les indices que nous considérons comme globaux, c'est-à-dire relatifs à l'ensemble de la scène et non spécifiques à un quelconque objet présent dans la scène. La sémantique de ce type d'indice portera donc sur l'image complète et non sur une partie de celle-ci. Il en résulte des indices fournissant une information purement temporelle.

Une scène extraite d'une séquence vidéo peut être caractérisée par deux aspects complémentaires : l'environnement qu'elle représente et la manière dont elle est acquise (par un capteur, le plus souvent une caméra). Les indices que nous allons présenter ici sont donc liés soit au capteur, soit à l'environnement.

2.1. *Indices liés au capteur*

Dans le cas d'une séquence vidéo acquise à l'aide d'une seule caméra (sans effet de post-production), statique de surcroît, la seule information liée au capteur et pouvant être extraite de la séquence vidéo concerne les paramètres intrinsèques de la caméra (focale par exemple). Nous ne nous intéressons pas ici à ce cas d'école et nous nous focalisons plutôt sur le cas plus complexe (mais aussi plus riche pour les problèmes d'indexation) de séquences vidéo obtenues après montage de plusieurs plans fournis par des caméras en mouvement. Dans ce contexte, les informations à extraire de la séquence vidéo sont liées soit au montage (ou post-production), soit aux caméras considérées individuellement.

Le montage consiste à assembler les différents plans au sein d'une même séquence vidéo. Nous rappelons qu'un plan est une succession d'images obtenues par une acquisition unique et continue à partir d'une caméra. Les plans sont accolés les uns aux autres et séparés par des transitions. Il existe deux types de transition : les transitions brusques et les transitions progressives. Les transitions brusques (*cuts*), comme leur nom l'indique, permettent de relier directement deux plans

successifs A et B : la dernière image du premier plan A est directement suivie par la première image du second plan B. La séquence est donc de la forme $A_1 \ldots A_n B_1 \ldots B_n$. Les transitions progressives consistent pour leur part en l'insertion d'un certain nombre t d'images intermédiaires I entre les plans A et B. Ces images intermédiaires sont obtenues par combinaison de deux images : la dernière image du premier plan (A_n) et la première image du second plan (B_1), selon la formule :

$$I_k = (1 - k/t)\, A_n + (k/t)\, B_1$$

pour la $k^{ième}$ image intermédiaire, avec k variant de 0 à t.

Différents types de transitions progressives existent et les plus connus sont le volet (*wipe*) et le fondu (*fade*). Les changements progressifs sont généralement plus difficiles à détecter que les changements brusques. Pour un panorama assez complet des techniques de détection des changements de plan dans des séquences vidéo non compressées, on pourra se référer à (Lefèvre *et al.*, 2003). Dans le cas d'une transition brusque, il sera possible d'extraire un indice (l'instant du changement) tandis que deux indices (le début et la fin de la transition) pourront être extraits lors de changements progressifs.

En ce qui concerne les indices directement relatifs au capteur (la caméra), il est possible de faire appel aux méthodes de caractérisation de caméra (Günsel *et al.*, 1998). Ainsi, outre la focale, différents paramètres peuvent être obtenus : déplacement ou translation (dans les axes X, Y, Z), rotation (dans les axes X,Y,Z), zoom, etc. A partir des informations précédentes, il est possible d'extraire des indices temporels liés à l'activité de la caméra : translation de l'instant t_i à t_j, rotation de t_k à t_l, zoom de t_m à t_n, etc. Le degré de précision des indices dépendra des besoins et on peut obtenir des informations du type $\{R_x = 0 ; R_y = 0 ; R_z = -90\}$ pour une rotation de -90° sur l'axe Z de t_i à t_j, puis $\{T_x = 1{,}4 ; T_y = 0 ; T_z = 0\}$ pour une translation de 1,4 mètres sur l'axe X de de t_j à t_k, etc.

2.2. *Indices liés à l'environnement*

Le second type d'indice global abordé ici est lié non pas au capteur mais à l'environnement. Dans une séquence vidéo, différents changements peuvent intervenir au sein même de la scène et les indices temporels qui y sont associés doivent donc être également extraits. Dans le cas d'une scène extérieure (mais aussi parfois d'une scène intérieure), les changements d'éclairage (ou d'illumination, de luminosité) ont suffisamment d'importance sur le contenu de la séquence et son exploitation pour être considérés comme des indices à extraire : images trop sombres (par exemple scène nocturne) ou trop saturées (par exemple soleil rasant) pour pouvoir être analysées correctement. Pour mesurer un changement d'éclairage, il est nécessaire de comparer les images de la séquence, le plus souvent de manière globale. Une mesure telle la différence des histogrammes peut être utilisée sur des images couleur ou niveaux de gris. Dans ce dernier cas, cette différence s'écrit :

$$D(I_k, I_l) = \sum_{n=0}^{255} |H_k(n) - H_l(n)|$$

avec $H_k(n)$ le nombre de pixels de I_k ayant pour niveau de gris la valeur n.

Il est aussi possible d'analyser directement une image (au travers de son histogramme par exemple) pour déterminer si elle est trop sombre, trop saturée, etc. Dans tous les cas, les indices liés à ces changements d'illumination peuvent être extraits.

Au cours d'une séquence vidéo, différentes scènes peuvent être représentées successivement. Il est fréquent d'observer dans un film une scène intérieure (dans une maison), composée de plusieurs plans successifs, puis une scène extérieure (dans un jardin), composée elle aussi de plusieurs plans. Dans ce contexte, il est intéressant d'extraire l'information liée à ce type de changement, dit changement de scène. Détecter un changement de scène s'avère en général plus délicat que détecter un changement de plan. Par analogie à la linguistique, on pourrait assimiler la scène à une notion sémantique et le plan à une notion syntaxique : les premières sont plus complexes à appréhender que les secondes. Les méthodes permettant la détection des changements de scène se basent sur les caractéristiques communes des différents plans d'une même scène : généralement, le décor reste identique, et il n'est pas rare d'observer des images successives aux histogrammes ou aux couleurs assez proches.

Nous avons détaillé ici les principaux indices globaux (relatifs à la scène) qui peuvent être extraits d'une séquence vidéo. Ces indices fournissent généralement une information succincte (instant d'un changement de plan, d'illumination, de scène, etc.). A partir d'un découpage en plans ou en scènes, il est également possible d'extraire des indices visuels tels que des images-clé (*keyframe*) (Hanjalic *et al.*, 1999). Une image-clé est une image qui a pour but de résumer le plan ou la scène (la succession de plans représentant le même environnement). Pour construire une image-clé, deux solutions alternatives existent : la première consiste à choisir une image présente dans le plan ou la scène (l'image la plus représentative), tandis que la seconde considère que l'image-clé est obtenue par combinaison (Moyen Âge, par exemple) des différentes images du plan ou de la scène. Les images-clé peuvent donc aussi être utilisées comme des indices temporels (associés aux extrémités du plan ou de la scène), même s'ils fournissent une information spatiale.

Dans les prochains paragraphes, nous allons présenter les indices relatifs aux objets présents dans la scène (donc de nature plus locale).

3. Indices relatifs aux objets

Une séquence vidéo est généralement caractérisée par la scène qu'elle représente mais aussi par les objets présents dans la scène. Ces objets se déplacent au sein de leur environnement (notion commune à tous les objets, et étudiée dans la partie

précédente), et interagissent soit entre eux, soit avec l'environnement. A partir de ce constat, nous avons décidé de nous intéresser aux indices relatifs aux objets en considérant successivement ceux liés à leurs mouvements (ou déplacements) et ceux liés à leurs activités (ou actions). Les premiers peuvent être assimilés à une information de bas niveau, tandis que les seconds représentent plutôt une information de haut niveau.

3.1. *Indices liés au mouvement des objets*

Le moyen le plus courant d'analyser les objets présents dans une séquence vidéo (et donc d'en extraire des indices) consiste à étudier leur position au cours du temps, donc à caractériser leur mouvement. Ce problème, connu sous le nom de suivi d'objet, fait l'objet de nombreux travaux (Moeslund *et al.*, 2001 ; Zhong *et al.*, 2000), généralement adaptés à un type d'objet particulier : objet rigide (véhicule), non rigide (personne, visage (Yang *et al.*, 2002)), déformable (nuage, organe), etc. Le suivi d'objet est généralement formulé de la façon suivante : connaissant la position d'un objet O dans les trames précédentes, on souhaite déterminer sa position dans l'image courante. Pour rechercher l'objet dans l'image courante, l'approche la plus utilisée requiert l'identification des caractéristiques de l'objet (forme, couleur, texture, etc.) puis la localisation de ces caractéristiques dans l'image à analyser. Ces caractéristiques peuvent être obtenues après une étape d'apprentissage réalisée hors ligne et basée sur différentes représentations de l'objet à suivre. Il est également possible de rechercher dans l'image courante les caractéristiques les plus similaires à celles trouvées dans l'image précédente. Quelle que soit la méthode employée, le suivi d'objet fournit à chaque instant la position des différents objets, et il est donc facile d'obtenir des informations complémentaires telles que leur déplacement, leur vitesse ou encore leur accélération.

Dans une séquence vidéo, d'autres indices de bas niveau relatifs aux objets peuvent également être extraits. Ainsi, l'apparition d'un nouvel objet dans la scène, qu'il soit quelconque ou prédéfini, fournit une information complémentaire au suivi d'objet. De même, il est possible de localiser les instants où les objets disparaissent de la scène. Cette disparition peut être de deux natures différentes : définitive lorsque l'objet quitte réellement la scène ou provisoire lorsque l'objet est par exemple caché par un autre objet. On parle dans ce dernier cas de phénomène d'occlusion. Chacune de ces différentes situations peut fournir des indices temporels de bas niveau permettant de caractériser le contenu de la séquence vidéo. D'autres techniques plus évoluées ont pour but l'extraction d'indices de plus haut niveau.

3.2. *Indices liés aux actions des objets*

Outre les informations liées au mouvement ou à l'apparition/disparition des objets dans la scène, les techniques d'analyse vidéo peuvent fournir des informations de plus haut niveau, relatives aux activités des objets dans la scène (Haritaoglu *et al.*, 2000 ; Stauffer et Grimson, 2000). Une des actions les plus simples à identifier parmi celles effectuées par un objet dans une scène est un contact avec un autre objet. Si le suivi d'objet fournit la forme des différents objets à chaque instant, et selon certaines hypothèses sur la profondeur des objets dans le champ de vision, on peut assimiler l'intersection de deux formes à un contact entre les deux objets qu'elles représentent. Certaines techniques fournissent des informations plus riches qu'un simple contact entre deux objets : passage d'un objet d'une personne à une autre, dépôt/prise d'un objet par une personne, etc. Dans ce cas, il est nécessaire d'identifier parmi les objets suivis lesquels sont acteurs (les personnes) et lesquels sont objets (valise par exemple).

Les activités des objets ne se limitent pas à des contacts ou des actions avec les autres objets. En effet, les objets peuvent également interagirent avec leur environnement. Ce type d'activité, pour être analysé correctement, nécessite une étude approfondie de l'environnement (généralement l'arrière-plan de la scène). L'obtention d'un modèle (par exemple 3-D) de l'environnement puis la localisation des objets non plus dans l'image mais au sein de l'environnement (*via* son modèle) fournit une information relative à l'interaction d'un objet avec son environnement. Cette information peut être utilisée pour extraire des indices temporels tels que la présence d'un objet dans une zone donnée de la scène.

Les méthodes présentées précédemment, prises individuellement, permettent de fournir des indices temporels (de plus ou moins haut niveau) relatifs au contenu global ou local de la séquence vidéo. En combinant ces différentes techniques, il est aussi possible de définir des événements à forte valeur sémantique : parking complet ou altercation entre 2 personnes dans une scène sous vidéosurveillance (Hu *et al.*, 2004), inscription d'un but dans un match de football, etc. La détection d'événements est un problème difficile à résoudre et les solutions proposées intègrent généralement, au sein d'une architecture globale, les techniques mentionnées auparavant pour des problèmes particuliers. La complexité de ce problème n'a pas permis pour l'instant d'obtenir de solutions génériques : les algorithmes sont aujourd'hui élaborés de façon *ad hoc* pour détecter tel événement particulier. Ces événements, à forte valeur sémantique, peuvent très logiquement être associés avec des indices temporels, caractérisés par un intérêt important pour l'utilisateur et une grande richesse d'expression. Néanmoins, ce problème reste ouvert dans le domaine de l'analyse vidéo.

Un récapitulatif des différents indices introduits dans cette section est présenté dans le tableau 1. En se basant sur ces indices, il est possible de construire des descripteurs représentant l'annotation (ou les métadonnées) des différents objets. Ces objets seront considérés, selon l'application, le contexte et les outils utilisés, à

différents niveaux de granularité. L'objectif est de permettre d'élaborer des requêtes sur un entrepôt construit à partir des données (semi-) struturées constituant les descripteurs. L'intérêt de cette approche est de permettre une exploitation des différentes dimensions de ces représentations multidimensionnelles (Chrisment et Sèdes, 2003).

Niveau de l'indice	Nature de l'indice	Problème d'analyse vidéo correspondant
Global (scène)	Capteur	Changement de plans
		Caractérisation de caméra
	Environnement	Changement d'illumination
		Changement de scène
	Combinaison	Image-clé
Local (objets)	Mouvement	Suivi d'objet
		Apparition/Disparition (Occlusion)
	Activités	Actions entre objets (contact)
		Interaction avec l'environnement
Combinaison		Détection d'événements

Tableau 1. *Taxonomie des indices temporels pouvant être extraits par des techniques d'analyse vidéo*

4. Bibliographie

Antani S., Kasturi R., Jain R., « A Survey on the Use of Pattern Recognition Methods for Abstraction », *Indexing and Retrieval of Images and Video, Pattern Recongition*, vol. 35, 2002, p. 945-965.

Correia P., Pereira F., « The Role of Analysis in Content-Based Video Coding and Indexing », *Signal Processing*, vol. 66, 1998, p. 126-142.

Chrisment C., Sèdes F., « Media Annotation », in *Multimedia Mining*, Kluwer Ac. Pub. ISBN 1-4020-7247-3, 2003, p. 197-208.

Günsel B., Tekalp A.M., van Beek P.J.L., « Content-Based Access to Video Objects: Temporal Segmentation, Visual Summarization, and Feature Extraction », *Signal Processing*, vol. 66, 1998, p. 261-280.

Haritaoglu I., Harwood D., Davis L., « W4: Realtime Surveillance of People and Their Activities », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 8, 2000, p. 809-830.

Hu W., Tan T., Wang L., Maybank S., « A Survey on Visual Surveillance of Object Motion and Behaviors », *IEEE Transactions on Systems*, Man and Cybernetics (Part C), vol. 34, n° 3, 2004, p. 334-352.

Hanjalic A., Zhang H., « An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis », *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, n° 8, 1999, p. 1280-1289.

Lefèvre S., Holler J., Vincent N., « A Study of Real-Time Segmentation of Uncompressed Video Sequences for Content-Based Search and Retrieval », *Real Time Imaging*, vol. 9, n° 3, 2003, p. 73-98.

Moeslund T.B., Granum E., « A Survey of Computer Vision-Based Human Motion Capture », *Computer Vision and Image Understanding*, vol. 81, n° 3, 2001, p. 231-268.

Stauffer C., Grimson E., « Learning Patterns of Activity Using Real-Time Tracking », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 8, 2000, p. 747-757.

Yang M.H., Kriegman D., Ahuja N., « Detecting faces in images: A survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, n° 1, 2002, p. 34-58.

Zhong Y., Jain A.K., Dubuisson-Jolly M.P., « Object tracking using deformable templates », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 5, 2000, p. 544-549.