Résumé de vidéo par détection de visage

Jean Emmanuel Viallet

France Télécom Recherche & Développement

Technopole Anticipa
2 Avenue Pierre Marzin, BP 40
22307 Lannion Cedex France
jeanemmanuel.viallet@francetelecom.com

Résumé
Les techniques de découpage en plans et la sélection d'images caractéristiques se fondent sur des indices de bas niveau indépendamment du contenu de l'image. Dans une image, les visages des personnes présentes est la première information regardée. L'indexation « visage » d'une vidéo permet de générer automatiquement des résumés « visage » en privilégiant dans le résumé les images caractéristiques où sont présentes des personnes.

Mots Clef
Vidéo, résumé, visage, détection, plan.

1 Introduction

1.1 Des résumés de nature différente
Les résumés de vidéos mettent l'accent soit sur la dimension émotive (bande annonce) soit sur la dimension informationnelle et plus couramment sur la dimension pragmatique. Ainsi quotidiennement, les plus grands tirages de la presse (la presse TV), résument, simplement, un film, à une image et plus généralement à une partie d'une image : un visage.
La dimension émotive ne respecte pas nécessairement ni la chronologie ni même le contenu en s'appuyant sur des images non présentes dans le document.
La dimension informationnelle devrait être comptable de la chronologie, de la durée et du contenu. Elle est utilisée par exemple dans les techniques de montage virtuel.
Un résumé peut se présenter sous la forme d'une vidéo [1] ou plus classiquement sous celle d'une mosaïque d'images fixes, cette représentation étant également celle qui est utilisé par les moteurs de recherche d'images fixes [2].

1.2 Un résumé fidèle basé sur le plan
Chaque plan identifié peut être représenté par une image fixe caractéristique qui résumerait ce plan. Le découpage en plan consiste à déterminer le lieu et la nature de la transition entre deux plans. De nombreux travaux ont permis de découper automatiquement une vidéo, en faisant appel à des algorithmes différents selon la nature de la transition entre plan.
Les techniques de découpage en plans et la sélection d'images caractéristique se fondent sur des indices de bas niveau (colorimétrie, mouvement) [3] et rien n'est connu du contenu de l'image caractéristique (présence/absence d'objets particuliers, de personnes).
Selon la durée, la nature et le rythme du montage, le nombre de plans peut être élevé et un résumé vidéo exhaustif (au moins une image par plan) important. On estime qu'il y entre 500 et 1000 plans/heure. Pour les vidéos que nous avons traités, le rythme est de 737 plans à l'heure. A ce rythme, une vidéo de 1h30 comporte plus de 1000 plans. Résumée en une seule image de format PAL, chaque plan aurait une taille de 25*19 pixels, la définition en dessous de laquelle il devient difficile d'identifier un visage.

1.3 Vers des résumés symboliques et compacts
Une séquence ou scène est une unité narrative, d'abstraction supérieure aux différents plans qu'elle regroupe. C'est une notion subjective qui varie selon le réalisateur, le monteur ou l'observateur.
Ainsi, un certain nombre d'images caractéristiques peuvent être jugées peu informatives (plans de coupe entre deux scènes) ou semblables (scène composée d'une succession de champ, contre-champ) et pouvant être éliminées du résumé.
A l'heure actuelle, on sait découper automatiquement en plan une vidéo, mais on ne sait pas découper en séquence. Le découpage en plan s'appuie sur des primitives bas niveau de l'image. Un découpage en séquence devrait s'appuyer sur des indices de haut niveau concernant le moment, le lieu, l'action, les protagonistes.

1.4 Un résumé compact très répandu :
Les programmes TV des journaux présentent sur une même page papier ou écran, les différentes émissions (éventuellement accompagnés d'un résumé image-visage) correspondant à la requête : quels sont les programmes du jour, dans telle tranche horaire, sur tel bouquet de chaînes? Par quelle image, l'émission est-elle résumée ?

En général par l'image d'une personne et plus particulièrement par un gros plan de cette personne.

Les moteurs de recherche qui fournissent des listes de résumés textuels ou des mosaïques d'images résumées par sous-échantillonnage s'appuient également sur ce mode de présentation par page qui facilite la consultation et la sélection par l'utilisateur.
Entre un résumé automatique systématique par plans et un résumé manuel où un opérateur sélectionne une seule image, nous proposons un résumé visage c'est à dire un résumé qui privilégie les images caractéristiques où sont présentes des personnes.
En s'appuyant sur ce critère de visage, on peut envisager de faire varier la taille du résumé selon les ressources disponibles.

2 Détection de visage

Il s'agit de classer les plans de la vidéo selon qu'ils présentent ou non des visages [4]. La recherche de visage peut s'effectuer soit sur l'ensemble du plan, soit sur une image caractéristique extraite du plan.

2.1 Détection de visage sur une image fixe

Le détecteur développé [5] détecte les visages de face et de profil (jusqu'à 60°) et fournit l'échelle et la position du visage.
Le détecteur détecte des visages en niveaux de gris. L'information de couleur chair permet de délimiter l'espace où sont recherchés les visages ce qui présente l'avantage d'accélérer le processus de détection, et l'inconvénient d'écarter les parties de l'image qui ne sont pas de teinte chair; ainsi le visage sur la photo sépia n'est pas détecté (Figure 2, $12^{ième}$ image).
Evalué sur une base de 13000 images fixes très variées collectées sur la toile (et non corrélées entre elles comme dans une vidéo) le taux de détection est de 75%, le taux de fausses alarmes est de l'ordre de 10^{-7} soit une fausse détection pour 250 images traitées. Le temps de traitement d'une image, en utilisant la couleur, est de l'ordre 1 seconde par image.
En ne traitant qu'une image caractéristique par plan, le temps de traitement par plan est minimal mais rien ne garantit qu'une personne présente dans ce plan sera détectée : la personne peut être absente de cette image ou dans une configuration telle qu'elle n'est pas détectée (cf. Figure 2, la $21^{iéme}$ image).

2.2 Détection de visage dans les vidéos.

Dans une vidéo, on dispose d'une information temporelle. Ainsi le mouvement peut être utilisé pour délimiter l'espace de recherche de visages. Cet indice est inopérant en cas de mouvement de caméra ou d'image fixe (à nouveau le cas de la photographie sépia).

Figure 1 : Détection : ensemble des 28 plans où des visages sont détectés Sur les 185 plans examinés, 50 comportent des visages, dont plusieurs semblables.

En parcourant toutes les images d'un plan jusqu'à la détection d'un visage, on favorise le processus de détection mais on pénalise le temps de traitement, surtout pour les plans où il n'y a pas de visage et qui sont testés du début à la fin.
Un tel temps de traitement ne permet pas un traitement en temps réel, c'est à dire égal à la durée de la vidéo.

Figure 2 Silence : ensemble des 22 plans où des visages présents ne sont pas détectés. Les principales raisons de la non-détection sont l'orientation, la taille, l'occultation et la colorimétrie du visage

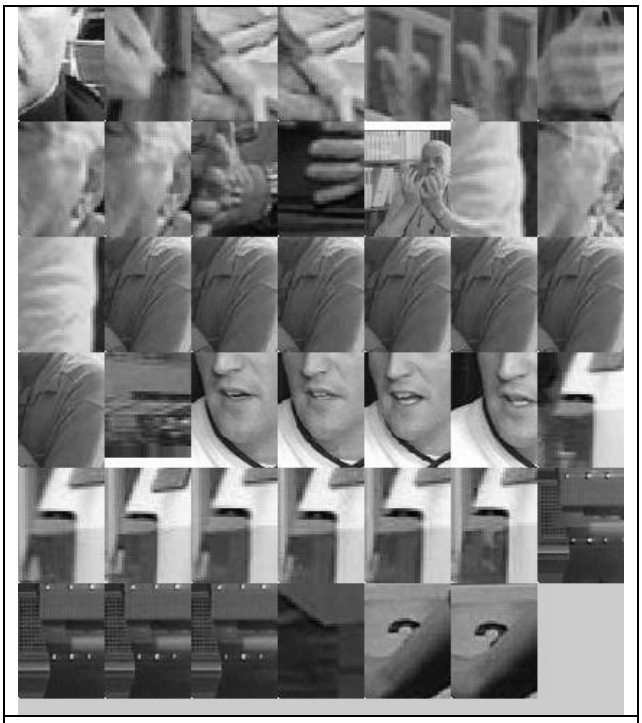

Figure 3 : Bruit : ensemble des 41 fausses alarmes sur 22337 images traitées, soit une fausse alarme pour 544 images. Dans une vidéo, les images fortement corrélés : c'est également le cas des fausses alarmes.

2.3 Détection de visages par échantillonnage

Pour les vidéos, un compromis doit être trouvé entre le temps de traitement et le nombre de visages détectés.

Dans les vidéos traitées, la durée d'un plan est en moyenne de 109 images soit de l'ordre de 4 secondes. Un traitement temps réel de la vidéo permet de traiter jusqu'à quatre images du plan avec le détecteur de visages.

En supposant connu le découpage en plan, un compromis rendement de détection - temps de calcul compatible avec un traitement temps réel semble obtenu (Figure 4) en échantillonnant régulièrement chaque plan avec au plus 4 images échantillons régulièrement espacés dans le plan. L'échantillonnage d'un plan est arrêté dès qu'une image échantillon avec des visages est obtenue. En moyenne, 3,65 images sont traitées pour 4 échantillons par plan.

Le rendement de détection est le rapport entre le nombre de détection pour E échantillons et le nombre de détection en explorant l'ensemble du plan. En passant de 3 à 4 échantillons par plan, le nombre de détections est inchangé : le nombre de visages détectés augmente et le nombre de fausses alarmes diminue. Les fausses alarmes ne peuvent être identifiées en tant que telles qu'après un dépouillement manuel de toutes les détections. La faible stabilité temporelle des fausses alarmes est une piste pour les identifiées automatiquement [6].

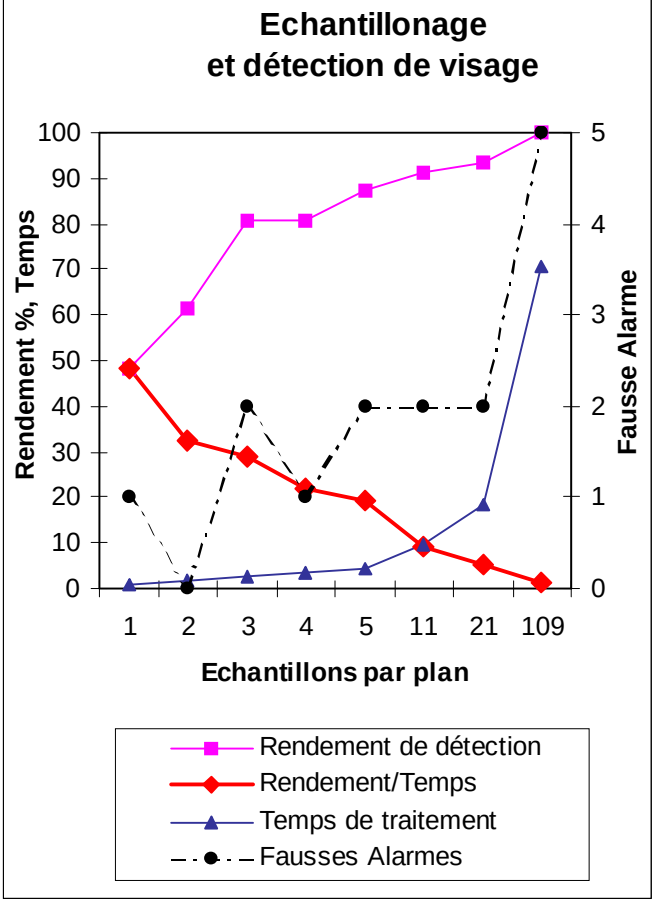

Figure 4 : Echantillonnage de 7 vidéos (14'53") soit 185 plans et 22337 images (109 images/plan , 737 plans/heure en moyenne). Pour 3 échantillons/plan : 512 images traitées, 23 visages détectés, 17 non détectés, .2 fausses alarmes.

3. Résumés de vidéos

Figure 5: Résumé de l'une des 7 vidéos traitées. Le

résumé « plan » est constitué des 26 images caractéristiques extraites des plans identifiés.

Figure 6 : Résumé « visage ». Les visages détectés sont indiqués par un cadre blanc. Plusieurs visages peuvent être détectés sur une image.

Figure 7 : Le résumé « visage » peut être enrichi en intercalant des images caractéristiques issues du découpage en plan.

Issu du découpage de la vidéo en plans , le résumé « plan » (Figure 5) comporte un nombre d'image égal au nombre de plan. Chacun des plans (et des images associées) a, a priori, la même importance.

Après le découpage en plan, le résumé « visage » (Figure 6), plus compact, ne retient que les images où des visages sont détectés. A la limite, ce résumé pourrait n'être constitué que d'une seule image. L'image correspondant au premier visage détecté limite le temps de traitement.

Un traitement plus long consiste à sélectionner une image du visage qui apparaît dans le plan le plus long.

Un traitement plus complexe consiste à sélectionner une image correspondant au visage qui apparaît dans le plus grand nombre de plans (par exemple, les deux premiers visages de la figure 6). Pour cela, il faut s'assurer qu'il s'agit du même visage [7]. Cela peut être relativement simple quand les images présentent peu de différence (les deux premières images de la figure 6) et plus difficile pour les images 3 et 4 de la figure 6. Le décor est différent, la pose du visage a changé et de plus le visage n'est pas vu de face ; or la plupart des techniques de reconnaissance de visage ne fonctionnent correctement que sur des visages vu de face [8].

Entre le résumé constitué d'une seule image, et le résumé complet, différents résumés de longueur croissante peuvent être obtenus en enrichissant le résumé « visage » avec des images extraites du découpage en plan (Figure 7).

4. Conclusion

La technique de détection de visage permet de diminuer la taille des résumés des vidéos et ne retenant que les images où sont présents les visages. Les résumés obtenus sont compacts car tous les visages ne sont détectés et que le nombre de fausses alarmes est bas. On constate que de nombreuses images du résumé sont semblables et pourraient être éliminées.

Bibliographie

[1] I. Yahiaoui, B. Merialdo, B. Huet, Résumés automatique de sequences vidéo, *CORESA'2000*, Poitiers, 19-20 octobre 2000

[2] J-Y Chen, C. Taskiran, A. Albiol, E. J. Delp and C. A. Bouman, "*ViBE:* A Video Indexing and Browsing Environment," *Proceedings of the SPIE Conference on Multimedia Storage and Archiving Systems IV,* September 20-22, 1999, Boston, vol. 3846, pp. 148-164.

[3] C.H. Demarty and S. Beucher., Efficient morphological algorithms for video indexing, *Content-Based and Multimedia Indexing, CBMI'99*, october 1999.

[4] Y. Chan, S.H. Lin, Y.P. Tan, S.Y. Kung, Video Shot Classification Using Human Faces, *ICIP(C)*, 1996, pp. 843-846.

[5] R. Féraud, O.J. Bernier, J.E. Viallet and M. Collobert , A fast and accurate face detector based on neural networks , *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, pp. 42-53, 2001.

[6]H. Wang, H. S. Stone, and S.-F. Chang, "FaceTrack: Tracking and Summarizing Faces from Compressed Video", *SPIE Multimedia Storage and Archiving Systems IV*, 19-22 Sept, 1999, Boston, MA.

[7]S. Eickeler, F. Wallhoff, U. Iurgel, and G. Rigoll. Content-Based Indexing of Images and Video Using Face Detection and Recognition Methods. *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City , Utah, May 2001.

[8] S. Satoh, Comparative Evaluation of Face Sequence Matching for Content-based Video Access, *Proc. of Int'l Conf. on Automatic Face and Gesture Recognition (FG2000)*, pp. 163-168, 2000.