Résumé de vidéo à partir d'un modèle d'attention visuelle

Sophie Marat, Mickaël Guironnet, Denis Pellerin

Résumé de vidéo à partir d'un modèle d'attention visuelle

Sophie Marat, Mickäel Guironnet, Denis Pellerin

Grenoble Images Parole Signal Automatique (GIPSA-lab) (ex LIS)
46 avenue Félix Viallet, 38031 Grenoble, France
sophie.marat@lis.inpg.fr, denis.pellerin@lis.inpg.fr

Résumé − Ce papier présente une méthode d'élaboration de résumé de vidéo, utilisant un modèle d'attention visuelle. Le modèle d'attention fournit des cartes de saillance qui mettent en évidence les zones des images contenant le plus d'information et donc susceptibles d'attirer le regard humain. Les cartes de saillance sont ensuite utilisées pour détecter les changements dans les images de la vidéo afin de permettre la sélection des images clés. La comparaison du résumé de vidéo obtenu avec un résumé de référence montre l'efficacité de la méthode proposée.

Abstract − This paper presents a method of video summarization based on a visual attention model. This model gives saliency maps which highlight area of frames containing more information and which attract human gaze. These saliency maps are used to detect changes on frames during the video which make it possible to select keyframes. A comparison between the summary and a reference summary shows the efficiency of our method.

1 Introduction

Avec l'augmentation continue du volume de données audiovisuelles, la recherche d'information et de documents pertinents devient un véritable défi. Une solution possible est la création automatique de résumé de vidéo par sélection d'images clés (résumé statique), qui facilite la navigation dans les bases de vidéos et l'extraction des évènements marquants. Les approches les plus courantes de résumé de vidéo utilisent des informations de bas niveau (couleur, texture...)[1, 2] qui sont peu représentatives du contenu sémantique de la vidéo.

Nous proposons dans cet article une méthode de résumé statique qui s'appuie sur une information de plus haut niveau extraite par un modèle d'attention visuelle. Ce modèle d'attention fournit des cartes de saillance qui mettent en évidence les zones des images contenant le plus d'information et donc susceptibles d'attirer le regard humain. Ces cartes de saillance sont utilisées pour détecter les changements dans les images de la vidéo afin de permettre la sélection des images clés.

Le modèle d'attention visuelle est décrit dans la section 2. La méthode de résumé est exposée dans la section 3. La méthode d'évaluation et les résultats sont présentés respectivement dans les sections 4 et 5.

2 Modèle d'attention visuelle

Le modèle d'attention visuelle le plus connu est celui proposé par Itti et Koch [3]. C'est une approche ascendante (« bottom-up ») qui prend en compte un grand nombre de caractéristiques visuelles telles que la couleur, l'intensité et l'orientation et qui génère une carte de saillance par image. Une carte de saillance est une image en niveau de gris où les zones claires correspondent aux régions qui vont le plus attirer le regard de l'observateur. Ce modèle statique, qui considère les images une par une, a été amélioré récemment en y intégrant le mouvement [4].

Ma et al. [5] ont proposé aussi un modèle d'attention. Ce modèle utilise un grand nombre de paramètres comme la saillance statique, la saillance du mouvement, le mouvement de caméra, la reconnaissance des visages, la saillance audio, les mots clés et les sujets clés. Un modèle global d'attention est construit à partir des attentions visuelle, audio et linguistique. Ce modèle d'attention est principalement utilisé pour du « video skimming », résumé qui se présente sous la forme d'une bande annonce. Un résumé statique peut aussi en être déduit mais ce modèle d'attention est trop complexe pour l'application au résumé statique seul.

Nous avons choisi d'utiliser le modèle d'attention que nous avons développé récemment et qui est décrit plus en détails dans [6]. Il prend en compte le mouvement et est plus simple que ceux proposés par Ma et al. [5] et Itti et al. [4]. Ce modèle extrait les caractéristiques de fréquence et d'orientation alors que celui proposé par Itti et al. considère en plus l'intensité et la couleur. L'originalité de notre modèle réside aussi dans la modélisation de la rétine au sens fonctionnel. Il associe deux voies parallèles (figure 1). La voie statique, d'inspiration biologique, extrait les régions texturées et contrastées de l'image. Elle comprend un filtre rétinien, un banc de filtres de Gabor et des interactions entre les réponses des filtres. Elle est issue des travaux précédents de Chauvin et al. [7]. La voie dynamique est utilisée pour détecter les objets en mouvement par une estimation et une compensation du mouvement de caméra ainsi qu'une différence des images compensées. Ce modèle donne, après un filtrage temporel et une normalisation, une carte de saillance pour chaque voie. Un exemple de cartes de saillance statiques est donné figure 2b.

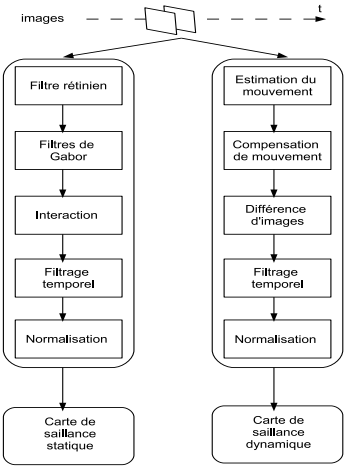

FIG. 1 – Principe du modèle d'attention

3 Méthode de résumé de vidéo

Une méthode de résumé de vidéo à partir d'un modèle d'attention a été proposé par Ma et al. [5]. Elle consiste à convertir la succession des cartes de saillance en une courbe d'attention et à sélectionner les images clés par détection des maxima sur cette courbe. La courbe d'attention est obtenue en remplaçant chaque carte de saillance par la moyenne de ses niveaux de gris. Les maxima de cette courbe correspondent aux images les plus saillantes c'est à dire les images les plus contrastées et texturées. Cette méthode présente l'inconvénient de conserver des images voisines temporellement qui sont susceptibles d'avoir un contenu similaire.

3.1 Principe de la méthode

Nous proposons une méthode de résumé de vidéo qui s'appuie sur le modèle d'attention présenté dans la figure 1. Cette méthode est composée de trois étapes : le choix du type de cartes de saillance à utiliser (statiques ou dynamiques), la sélection des images clés, et enfin l'élimination des images redondantes.

3.1.1 Choix du type de cartes de saillance

Cette étape correspond au choix de la catégorie de carte de saillance la plus appropriée pour chaque plan de la vidéo à résumer. Les cartes statiques mettent en évidence les objets texturés et contrastés et sont adaptées pour décrire les plans avec peu de mouvement. À l'opposé, les cartes dynamiques mettent en évidence les objets en mouvement et sont adaptées pour les plans avec un mouvement important de caméra ou des objets. Une courbe d'attention (comme définie par Ma et al.) est calculée pour chaque type de carte de saillance. La courbe d'attention dont l'écart-type des amplitudes est le plus grand correspond au type de carte donnant le plus d'information et qui doit donc être choisi. Nous avons vérifié expérimentalement que ce critère de choix apportait les meilleurs résultats pour le résumé final. En effet les résumés obtenus en utilisant les cartes de saillance appropriées retiennent plus d'évènements différents et moins d'images redondantes.

3.1.2 Sélection des images clés

Pour la seconde étape, l'objectif est de construire un résumé avec un nombre réduit d'images, suffisamment différentes pour bien couvrir le contenu de la vidéo. Cette étape repose sur une courbe de *variation* d'attention à partir de laquelle sont sélectionnées les images clés. Afin de mettre en évidence les changements dans la vidéo, une différence D_k entre cartes de saillance est calculée : $D_k = |M_k - M_{k-i}|$ où M_k est la carte de saillance associée à l'image k et i est le paramètre qui définit l'écart entre la carte courante et une carte calculée précédemment. Des tests ont montré que choisir i=10 permet d'avoir des images suffisamment différentes pour observer le changement et pour prendre en compte une variation rapide des cartes de saillance (les cartes de saillance s'éclaircissent au cours du temps s'il y a changement et s'assombrissent dans le cas contraire). La courbe de variation d'attention est alors obtenue en calculant la moyenne des niveaux de gris sur chaque différence de cartes de saillance.

La sélection d'images clés est faite en détectant sur la courbe de variation d'attention (exemple figure 2), un accroissement significatif par rapport au voisinage précédent. Pour cela un seuil adaptatif est utilisé. Ce seuil est donné par l'expression classique $\mu + 3\sigma$ où μ représente la moyenne glissante sur une fenêtre temporelle et σ l'écart type sur cette même fenêtre. La figure 2 montre que deux images de contenu similaire (ex : images 200 et 210) correspondent à des valeurs faibles sur la courbe de variation d'attention alors que les phases de changement (ex : 345 à 365) correspondent à des valeurs croissantes. Le seuil adaptatif permet de prendre en compte le passé et de déterminer si l'image diffère significativement des précédentes, et donc de retenir des images au niveau des transitions entre les différentes phases. Une seule image est retenue pour le résumé à chaque fois que la courbe de variation dépasse le seuil, même s'il est dépassé pour plusieurs images consécutives.

3.1.3 Élimination des images redondantes

Un résumé a été obtenu pour chaque plan. Ce résumé est fait en s'efforçant de ne pas manquer d'évènements dans la vidéo et en ne sélectionnant pas d'images voisines, cependant des images ayant un contenu similaire peuvent encore être retenues (elles sont cependant moins nombreuses qu'avec la méthode de Ma et al.). Un post-traitement est proposé afin d'éliminer les images redondantes (figure 3). Ce traitement considère la carte de saillance statique associée à chaque image. Une comparaison entre la carte de saillance de l'image courante et celle de l'image précédemment retenue est calculée. Si cette valeur est en dessous d'un seuil, les deux images sont trop proches et seule celle avec la valeur sur la courbe d'attention la plus élevée est retenue. Des tests ont permis de choisir le seuil égal à 0,11 et de s'assurer de sa robustesse.

Un exemple de résumé produit par cette méthode est donné à la figure 4, il prend bien en compte les différents évènements du plan. Cette méthode est utilisée pour les plans longs et une adaptation a été faite pour les plans courts.

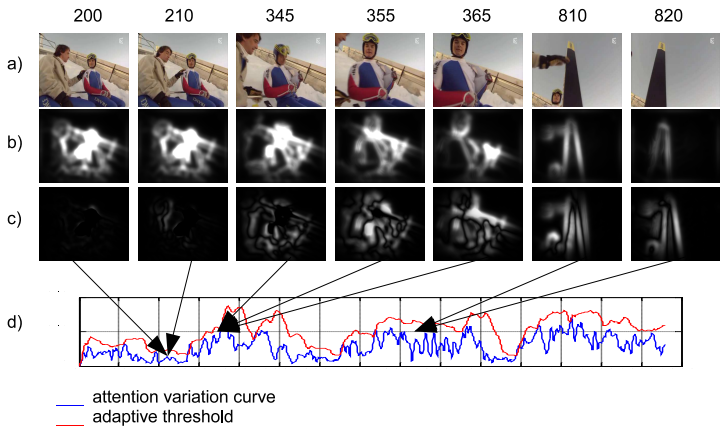

FIG. 2 – exemple de cartes de saillance pour un plan de l'émission éducative. a) images de la vidéo, b) cartes de saillance statiques, c) différence de cartes de saillance, d) courbe de variation d'attention.

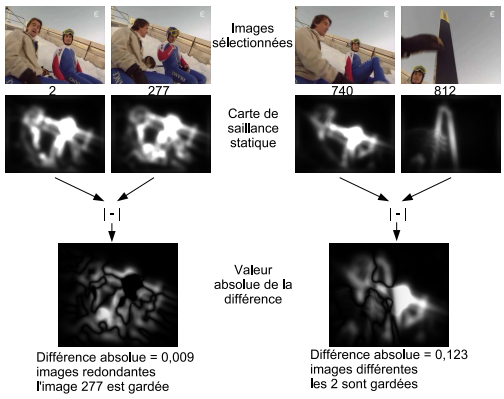

FIG. 3 – Application du post-traitement permettant d'éliminer les images redondantes. À gauche, les masques se ressemblent et leur différence est inférieure à 0,11 alors que dans l'exemple de droite, les masques sont différents et leur différence supérieure à 0,11.

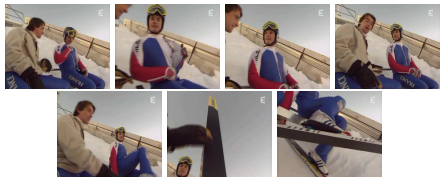

FIG. 4 – Résumé du plan décrit dans la figure 2a

3.2 Cas particulier des plans courts

Le cas particulier des plans courts est considéré ici. Un plan est une portion continue de vidéo avec ou sans mouvement de caméra. Nous définissons un plan court comme un plan de 100 images ou moins (i.e. inférieur à 4 secondes). La continuité temporelle des images de la vidéo fait que toutes les images d'un plan de 4 secondes ont un contenu proche. Une seule image clé est alors nécessaire pour le résumé. Cette image clé est sélectionnée en choisissant le maximum sur la courbe d'attention obtenue à partir des cartes de saillance statiques. En effet, dans un plan court le contenu varie peu, les cartes statiques sont donc plus appropriées que les cartes dynamiques.

4 Méthode d'évaluation du résumé

Il existe différentes méthodes d'évaluation des résumés de vidéo. Une possibilité est de demander à un sujet de choisir entre deux résumés [8]. Dans cet article l'évaluation du résumé obtenu se fait par comparaison avec une « vérité terrain » appelée résumé de référence.

Pour obtenir le résumé de référence, nous demandons à plusieurs sujets de regarder un plan d'une vidéo et d'en faire leur résumé idéal. Les instructions leur imposent de faire un résumé de 1 à 3 images pour chaque plan. Pour chaque plan le nombre N d'images clés pour le résumé de référence est obtenu en prenant le médian du nombre d'images retenues par tous les sujets. Selon le nombre d'images choisies par un sujet sur un plan, une pondération est donnée à chacune de ces images. Plus un sujet sélectionne d'images, plus leur poids est faible. Les images du résumé de référence sont sélectionnées en prenant les N images avec le poids le plus élevé.

4.1 Méthode de comparaison

La méthode de comparaison du résumé automatique, appelé résumé candidat, au résumé de référence se fait en 4 étapes (figure 5).

La première étape consiste à associer les images du résumé candidat à celles du résumé de référence. Chaque image du résumé de référence est associée à deux images (au maximum) du résumé de référence, l'image précédente et la suivante. Au cours de la deuxième étape l'image du résumé candidat est seulement associée à l'image du résumé de référence la plus proche temporellement. La troisième étape ne retient que les images du résumé candidat qui sont les plus proches de celles du résumé de référence. L'étape quatre compare les images du résumé candidat aux images associées dans le résumé de référence en faisant une comparaison de leur histogrammes couleur. Les images sont continues temporellement, il est donc improbable d'avoir deux images avec des histogrammes similaires et un contenu différent. Le descripteur utilisé ici est un histogramme couleur global et la distance entre histogrammes est obtenue par la norme L1.

TAB. 1 – Résultats pour 3 méthodes de résumé sur 3 vidéos (méthodes : n^o1 résumé aléatoire, n^o2 résumé sélectionnant 1 image au milieu de chaque plan, n^o3 résumé à partir du modèle d'attention).

résumé	**émission éducative**			**journal télévisé**			**série**		
	R	P	F1	R	P	F1	R	P	F1
n^o1	62 (15/24)	40 (15/37)	49.1	83 (46/55)	50 (46/91)	63.0	80 (24/30)	40 (24/59)	53.9
n^o2	50 (12/24)	60 (12/20)	54.5	63 (35/55)	83 (35/42)	72.1	73 (22/30)	78 (22/28)	75.8
n^o3	62 (15/24)	51 (15/29)	56.6	78 (43/55)	70 (43/61)	74.1	80 (24/30)	75 (24/32)	77.4

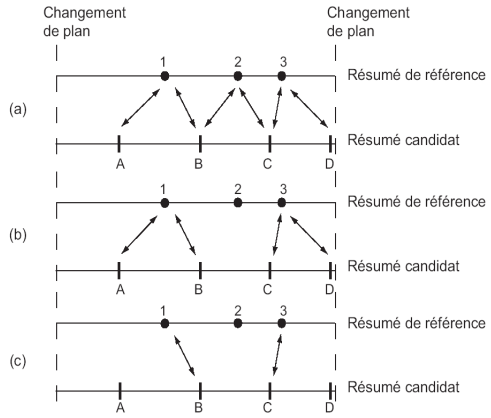

FIG. 5 – Trois premières étapes de la méthode de comparaison

5 Résultats expérimentaux

La méthode de résumé de vidéo a été testée sur 3 vidéos : une émission éducative, un journal télévisé et une série. Ces vidéos représentent 90 plans de 19 à 1468 images.

L'évaluation des résumés est faite en utilisant la méthode de comparaison décrite au paragraphe 4 et les critères de rappel R, précision P, et leur moyenne harmonique F1.

$$R = \frac{N_t}{N_r}, \quad P = \frac{N_t}{N_c}, \quad F1 = 2 \times \frac{R \times P}{R+P}$$

où N_t est le nombre d'images du résumé candidat qui correspondent à celles du résumé de référence, N_r le nombre d'images du résumé de référence, et N_c le nombre d'images du résumé candidat.

Deux autres méthodes de résumé (un résumé aléatoire et un résumé sélectionnant une image au milieu de chaque plan) sont aussi comparées au résumé de référence. Les résultats sont présentés dans le tableau 1. En prenant l'exemple du résumé utilisant le modèle d'attention (n^o3) et de la vidéo émission éducative, le rappel (respectivement, la précision) indique que sur les 24 (resp. 29) images du résumé de référence (resp. candidat) le résumé automatique en retrouve 15.

Le résumé utilisant le modèle d'attention donne les meilleurs résultats. Mais l'écart avec le résumé prenant une image clé au milieu de chaque plan n'est pas très grand. Ceci s'explique par le fait que ces vidéos présentent une majorité de plans courts (inférieur à 100 images), où les images choisies par les deux méthodes de résumé sont équivalentes. L'efficacité de la méthode de résumé à partir d'un modèle d'attention augmente avec la durée des plans. Une comparaison avec une méthode de résumé à partir du mouvement de caméra est présentée dans [9].

6 Conclusion et perspectives

Nous avons décrit une méthode de résumé de vidéo qui repose sur un modèle d'attention visuelle. Cette méthode utilise les cartes de saillance pour mettre en évidence les passages de la vidéo où des changements apparaissent. Elle a été testée sur 3 vidéos de taille et de contenu différents. Pour cela une méthode d'élaboration de résumé de référence a été proposée. Les résultats obtenus sont satisfaisants et s'améliorent avec la durée des plans.

L'approche précédente a été réalisée plan par plan dans la vidéo. Elle pourrait être généralisée au résumé global de vidéo de manière à diminuer la redondance et gagner en compacité du résumé.

Références

[1] Y. Li, T. Zhang, D. Tretter. *An overview of video abstraction techniques.* HPL-2001-191, 2001.

[2] H. J. Zhang, J. Wu, D. Zhong, S. W. Smoliar. *An integrated system for content-based video retrieval and browsing.* Pattern Recognition, vol. 30, N^o. 4, pp 643-658, 1997.

[3] L. Itti, C. Koch, E. Niebur. *A model of saliency-based visual attention for rapis scene analysis.* IEEE Trans. on pattern analysis and machine intelligence, vol. 20, pp. 1254-1259, 1998.

[4] L. Itti, N. Dhavale, F. Pighin. *Realistic avatar eye and head animation using a neurobiological model of visual attention.* SPIE 48^{th} Annual Int. Symp. on Optical Science and Technology, vol. 5200, pp. 64-78, 2003.

[5] Y. Ma, X. Hua, H. Zhang. *A generic framework of user attention model and its application in video summarization.* IEEE Trans. on multimedia, vol. 7, pp. 907-919, 2005.

[6] M. Guironnet, N. Guyader, D. Pellerin, P. Ladret. *Static and dynamic feature-based visual attention model : comparison with human judgement.* EUSIPCO 2005,Turkey.

[7] A. Chauvin, J. Herault, C. Marendaz, C. Peyrin. *Natural scene perception : visual attractors and image processing.* Connectionnist Models of Cognition and Perception, World scientific Press, 2002.

[8] S. Corchs, G. Ciocca, R. Schettini. *Video summarization using a neurodynamical model of visual attention.* IEEE 6^{th} Workshop on Multimedia Signal Processing, Sienne, Italy, pp. 71-74, 2004.

[9] M. Guironnet. *Méthodes de résumé de vidéo à partir d'information bas niveau, du mouvement de caméra ou de l'attention visuelle.* Thèse de doctorat de l'Université Joseph Fourier (Grenoble), 2006.