Segmentation vidéo interactive par zones quasi-plates

Jonathan Weber[1], Sébastien Lefèvre[2], Pierre Gançarski[1]

[1]Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection, Université de Strasbourg – CNRS
Pôle API, Bd Sébastien Brant, BP 10413, 67412 Illkirch Cedex, France

[2]Laboratoire de Recherche en Informatique et ses Applications de Vannes et Lorient, Université de Bretagne-Sud
Bât. ENSIbs, BP 573, 56017 Vannes Cedex, France

j.weber@unistra.fr, sebastien.lefevre@univ-ubs.fr, gancarski@unistra.fr

Résumé – Le volume des données vidéo ne cesse d'augmenter dans les bases de données personnelles et sur le Web. Pour exploiter ces données, une segmentation préalable est souvent nécessaire afin d'obtenir les objets d'intérêt à traiter ultérieurement. Cependant, la segmentation d'une séquence vidéo n'est pas unique et dépend des besoins de chaque utilisateur. Une segmentation personnalisée peut être réalisée en utilisant des méthodes interactives, mais seulement si leur temps de calcul reste raisonnable afin de permettre dans de bonnes conditions cette interactivité. Dans cet article, nous abordons le problème de la segmentation vidéo interactive et proposons une approche en deux étapes : 1) un traitement hors-ligne pour extraire automatiquement les zones quasi-plates à partir d'une séquence vidéo, et 2) un traitement en-ligne interactif destiné à assembler les zones quasi-plates afin de construire les objets d'intérêt. Notre approche est capable de faire face à de multiples objets, est robuste aux erreurs introduites par l'étape de présegmentation automatique et ne nécessite pas de réitérer l'ensemble du processus de segmentation à chaque correction des marqueurs par l'utilisateur.

Abstract – Video data is continuously increasing in personal databases and Web repositories. To exploit such data, a prior segmentation is often needed in order to extract the objects-of-interest to be further processed. However, the segmentation of a given video is often not unique and indeed depends on user needs. Personalized segmentation may be achieved using interactive methods but only if their computational cost stays reasonable to enable user interactivity. We address here the problem of interactive video segmentation and introduce a 2-step segmentation scheme: 1) offline processing to automatically extract quasi-flat zones from video data, and 2) online processing to interactively gather quasi-flat zones and build objects-of-interest. Our approach is able to deal with multiple objects, robust to errors introduced by the automatic presegmentation step, and does not require to perform again the whole segmentation process each time the user provides some feedback.

1 Introduction

Après l'augmentation des données de type texte et image, nous assistons actuellement à la prolifération des données de type vidéo. De nombreux traitements ou utilisations de ces données nécessitent une segmentation préalable afin d'obtenir les objets d'intérêt qui seront traités ultérieurement, citons par exemple la fouille vidéo [6]. La segmentation d'une vidéo n'est cependant pas unique et dépend des besoins des utilisateurs. Une telle segmentation personnalisée peut être obtenue par une méthode interactive, sous réserve que son efficacité soit suffisante pour permettre l'interactivité.

Nous proposons ici une méthode originale, interactive et efficace de segmentation vidéo basée sur deux étapes : 1) un traitement hors-ligne pour extraire des zones quasi-plates à partir de données vidéo et ainsi réduire le volume des données à traiter, et 2) un traitement en ligne pour fusionner interactivement les zones quasi-plates et construire les objets d'intérêt du point de vue de l'utilisateur. Quelques résultats préliminaires illustrent sa pertinence.

2 Zones quasi-plates

La définition originelle d'une zone quasi-plate (ZQP) notée α-$\mathcal{Z}$ est, pour un pixel p, l'ensemble connexe des pixels pouvant être atteints depuis p par (au moins) un chemin vérifiant la condition suivante : la différence entre les valeurs des pixels successifs du chemin est inférieure ou égale à un paramètre donné α. Par la suite, d'autres définitions de ZQP ont été proposées (cf. [5] pour une étude complète). Nous utiliserons ici le concept de (α,ω)-$\mathcal{Z}$ qui impose la contrainte supplémentaire suivante : la différence maximale entre les valeurs des pixels de la ZQP doit être inférieure ou égale au paramètre ω.

A la connaissance des auteurs, il n'existe pas d'extension du concept de ZQP aux données vidéo. L'extension la plus triviale est l'approche $3D$ qui consiste à représenter la séquence vidéo comme un cube et à considérer un voisinage spatio-temporel et non plus uniquement spatial. Toutefois, les dimensions spatiales et temporelles sont intrinsèquement différentes et les résultats fournis par une telle approche ne sont généralement pas pertinents (sous-

segmentation spatio-temporelle). Nous avons donc considéré ici une méthode $2D + t$ (cf. [7] pour plus de détails) qui produit dans un premier temps les ZQP en ne considérant que la dimension spatiale (i.e. indépendamment sur chaque trame). Les ZQP obtenues sont ensuite considérées comme des unités élémentaires, représentées au sein d'un graphe où chaque ZQP est un nœud valué (par exemple, par la valeur moyenne de ses pixels). La dimension temporelle est alors étudiée en reliant les ZQP de trames adjacentes et se chevauchant spatialement, ce lien étant matérialisé par des arêtes. Les arêtes sont valuées par une distance (ici la distance euclidienne) entre les valeurs des nœuds qu'elles relient. Les ZQP sont définies dans le graphe comme les plus grandes composantes connexes dont les arêtes ont une valeur inférieure ou égale à α et dont la différence entre les valeurs des nœuds ne dépasse pas ω. Nous pouvons également utiliser ce schéma pour une approche $t + 2D$ en traitant en premier lieu la dimension temporelle puis les dimensions spatiales. Pour réduire la sur-segmentation induite par les ZQP et assurer l'efficacité du processus de fusion, nous appliquons enfin, à l'instar de la démarche communément adoptée pour les images fixes, un processus de filtrage qui supprime les ZQP non-significatives, c'est-à-dire d'aire moyenne inférieure à un seuil a_m, en les fusionnant avec leurs ZQP voisines. Nous obtenons alors une sur-segmentation dont les régions représentent une réduction efficace de l'espace des données. La partition en ZQP vidéo peut être vue comme une approche créant des superpixels qui permettent ensuite de traiter la séquence vidéo plus rapidement.

3 Méthode de segmentation

L'approche de segmentation proposée ici (cf. figure 1) consiste en deux étapes : tandis que la première étape est effectuée hors-ligne et ne nécessite pas d'intervention de l'utilisateur, la seconde étape opère en ligne et est interactive.

L'étape hors-ligne est un pré-traitement qui a pour but de produire une première segmentation en ZQP de la séquence vidéo que ce soit par l'approche $2D+t$ ou $t+2D$ (cf. section 2 et [7]). Les ZQP obtenues sont représentées par un graphe d'adjacence de régions spatio-temporelles (GARST). L'essentiel du cout calculatoire de notre méthode est dû à cette étape, qui s'effectue toutefois hors-ligne.

La seconde étape opère en ligne et repose sur l'intervention de l'utilisateur. A l'aide de marqueurs éditables, celui-ci va raffiner interactivement la sur-segmentation initiale des ZQP jusqu'à produire une segmentation à sa convenance. Les marqueurs correspondent aux objets d'intérêt et au fond, et peuvent être définis dans des trames différentes. Ils sont donc définis spatialement par l'utilisateur qui les dessine sur une trame donnée, mais sont traités comme des marqueurs spatio-temporels. Pour cela, le GARST initial ou corrigé après plusieurs interactions est segmenté dans un cadre interactif, en utilisant un algorithme de segmentation inspiré du Seeded Region Growing (SRG) [1]. Chaque interaction consiste en différentes étapes : 1) les nœuds représentant des ZQP recouvertes par un marqueur sont considérés comme graines par le SRG et sont étiquetés *fusionné*, les autres nœuds étant étiquetés *non fusionné* ; 2) si des marqueurs différents recouvrent la même ZQP, cette ZQP est resegmentée selon ces marqueurs en utilisant la ligne de partage des eaux basée marqueurs [4] ; 3) chaque arête reliant un nœud *fusionné* à un nœud *non fusionné* est valuée par la distance mesurée entre les attributs de ces nœuds (par exemple, la distance euclidienne entre les couleurs moyennes de chaque nœud) ; 4) le GARST est simplifié itérativement (jusqu'à ce qu'il ne contienne plus de nœud *non fusionné*) en supprimant l'arête de plus petite valeur, en fusionnant les nœuds qu'elle reliait, et en mettant à jour les valeurs des arêtes du nouveau nœud *fusionné* ; 5) les nœuds relatifs au même marqueur sont fusionnés afin d'obtenir un unique nœud par marqueur ; 6) le GARST est utilisé pour produire un résultat de segmentation compréhensible par l'utilisateur. Si l'utilisateur n'est pas satisfait de la segmentation obtenue, il peut corriger ou supprimer les marqueurs existants mais aussi en ajouter de nouveaux. Le GARST est alors réinitialisé et le processus itératif est répété jusqu'à satisfaction de l'utilisateur.

4 Résultats

Pour illustrer le comportement de notre méthode, nous considérons ici la séquence de référence *Carphone* (381 trames de 176×144 pixels) et montrons un échantillon correspondant à la trame 186 en figure 2. Par souci de clarté, la figure présente des images fixes mais le processus de segmentation est spatio-temporel. Alors que la séquence vidéo originale (a) contenait 9 656 064 pixels, la segmentation initiale en ZQP (b) donne 35 560 régions. L'utilisateur doit dessiner un marqueur pour chaque objet d'intérêt et pour le fond, par exemple (c). Le résultat (d) est obtenu en ne marquant que quelques trames dans la vidéo. En effet, il est conseillé aux utilisateurs de ne marquer tout d'abord que la trame médiane afin de produire un premier résultat de segmentation. Celui-ci peut ensuite être corrigé par l'ajout, la modification, et/ou la suppression de marqueurs sur n'importe quelle trame.

La table 1 indique le temps de calcul nécessaire pour segmenter les 381 trames de la séquence *Carphone*. Le temps hors ligne représente le temps inhérent à la production des ZQP et à la création du GARST, alors que le temps en ligne représente le coût calculatoire d'une interaction avec l'utilisateur, i.e. une itération du processus de raffinement sur toute la séquence vidéo (un temps moyen par trame est aussi fourni). Nous avons comparé les temps de calcul de notre méthode (intitulée ZQP basée marqueurs – ZQPBM) avec une méthode récente de segmentation

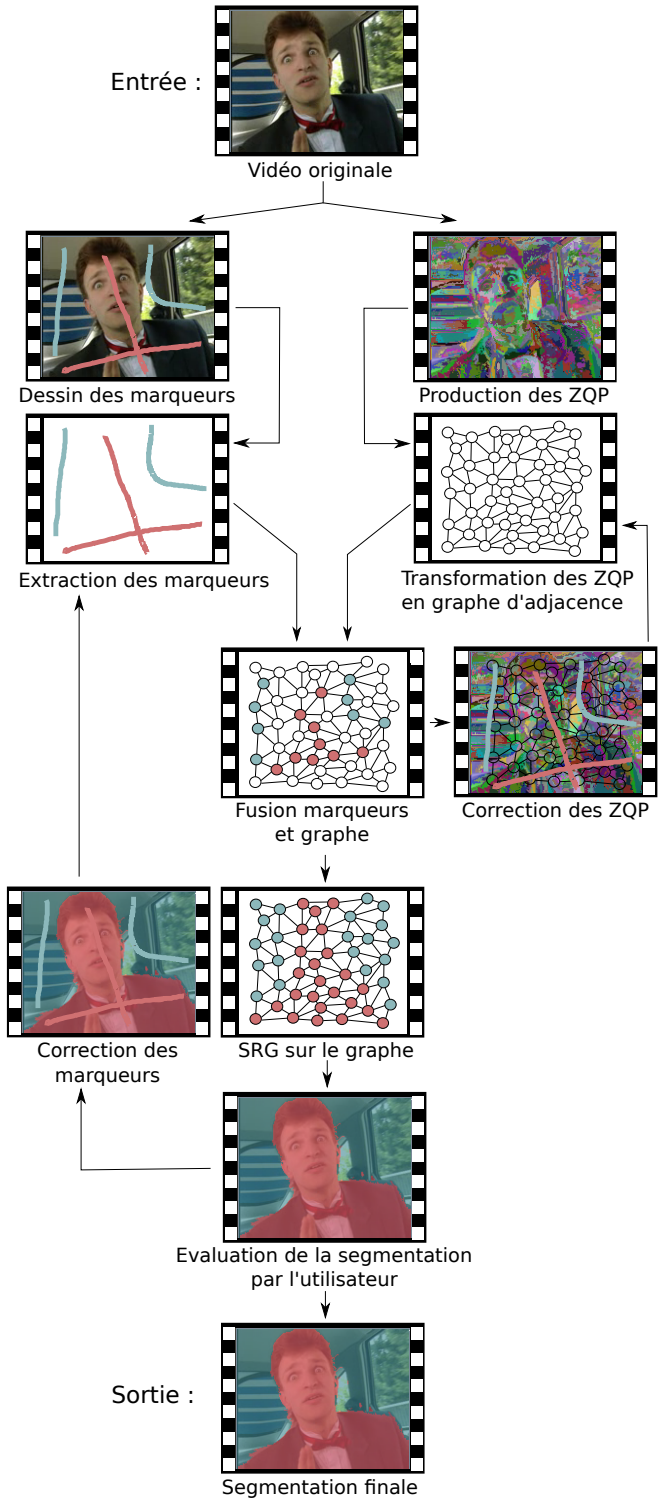

FIGURE 1 – Processus de segmentation vidéo interactive par les ZQP guidées par marqueurs

moprhologique interactive de séquences vidéos appelée *ligne de partage des eaux par propagation de marqueurs (LPEPPM)* [2] et avec deux méthodes de référence : *ligne de partage des eaux basée marqueurs (LPEBM)* [4] et

FIGURE 2 – Illustration du processus de segmentation avec les résultats obtenus sur la trame 186 : a) trame originale, b) sur-segmentation initiale, c) marqueurs, d) segmentation finale.

Seeded Region Growing (SRG) [1] étendues aux données vidéo (considérées ici comme des volumes $3D$). Seul le temps en ligne est mesuré ici : il décrit l'efficacité du processus interactif. A chaque itération, nous calculons donc la segmentation complète de la séquence vidéo. Cela rend la comparaison avec la LPEPPM injuste car cette méthode ne repose pas sur une segmentation globale mais trame par trame, l'objectif est de comparer le coût de notre approche à celui d'une approche plus complexe utilisant l'information de mouvement (ici par le mécanisme de propagation de marqueurs). Notre méthode est nettement plus rapide, dans sa partie en-ligne, que les trois autres approches grâce à son pré-traitement (segmentation en ZQP) effectué hors-ligne. Le faible temps de calcul qu'elle nécessite à chaque itération permet une interaction réelle avec l'utilisateur qui est confronté à un temps d'attente quasi-nul (inférieur à la seconde). Ces résultats valident notre choix d'utiliser des descripteurs très simples (couleur moyenne) pour les ZQP. Ce choix nous permet d'obtenir un temps de calcul très faible entre chaque interaction alors que l'utilisation de descripteurs plus complexes (comme pour la LPEPPM) nécessite un important temps de calcul.

Par ailleurs, nous avons également évalué la précision des résultats obtenus par notre méthode et celle d'autres approches. Pour cela, nous comparons notre approche à la LPEBM et au SRG. Nous avons choisi ces deux méthodes car nous utilisons la LPEBM pour resegmenter les ZQP remises en cause par les marqueurs ; et nous appliquons un SRG sur le GARST pour obtenir la segmentation désirée par l'utilisateur. Le but de cette comparaison est de montrer que notre approche, qui repose sur ces méthodes de base, donne de meilleurs résultats que leur application directe. Pour effectuer l'évaluation de la précision, nous utilisons un extrait de la séquence *carphone* de 80 trames pour lesquelles nous avons réalisé une vérité terrain. Nous évaluons les résultats en utilisant la valeur moyenne de

TABLE 1 – Evaluation de l'efficacité des méthodes de segmentation interactive sur la séquence *carphone* : temps de calcul (en secondes) des approches ZQPBM selon différents paramètres α et ω, LPEPPM, LPEBM et SRG.

Méthode	α, ω	Hors ligne	En ligne	(par trame)
ZQPBM 2D+t	10	44	0.52	($1.3x10^{-3}$)
	20	35	0.55	($1.4x10^{-3}$)
	30	38	0.50	($1.3x10^{-3}$)
	40	43	0.36	($9.6x10^{-4}$)
	50	46	0.32	($8.6x10^{-4}$)
ZQPBM t+2D	10	44	0.10	($2.8x10^{-4}$)
	20	32	0.12	($3.2x10^{-4}$)
	30	26	0.11	($3.0x10^{-4}$)
	40	26	0.10	($2.8x10^{-4}$)
	50	25	0.09	($2.6x10^{-4}$)
LPEPPM	–	3	132	(0.35)
LPEBM	–	3	27	(0.07)
SRG	–	0	56	(0.14)

l'indice de Jaccard qui a été utilisé dans l'évaluation de segmentation d'images [3]. Les résultats ont été obtenus par un jeu de marqueurs présent uniquement sur la trame médiane. Les résultats sont présentés dans la table 2 et montrent que notre approche de ZQPGM donne de bien meilleurs résultats que l'application direct du SRG sur les pixels. Ceci montre donc l'intérêt de construire une représentation intermédiaire des données, à l'aide de zones quasi-plates. De plus, nous observons que notre méthode est capable de donner des résultats légèrement meilleurs que ceux obtenus avec la LPEBM (pour un temps de calcul 50 fois moindre). Notons aussi que ces résultats sont obtenus par différentes combinaisons de paramètres α, ω et a_m, montrant ainsi la relative robustesse de notre méthode au réglage des paramètres. Cependant, la sélection des meilleurs paramètres pour notre approche est encore un problème ouvert.

TABLE 2 – Évaluation de la précision des segmentations obtenues sur un extrait de 80 trames de la séquence *carphone*. La précision correspond à l'indice de Jaccard moyen obtenu pour les 3 objets (homme, intérieur, extérieur), seule la trame médiane étant marquée. La ZQPBM est comparée aux approches SRG et LPEBM.

Méthode	(α, ω)	a_m	Ind. Jaccard
ZQPGM 2D+t	30	10	0.905
	50	50	0.910
	90	50	0.908
ZQPGM t+2D	20	60	**0.928**
	40	100	0.925
	100	70	0.919
SRG	-	-	0.548
LPEGM	-	-	0.897

5 Conclusion

Dans cet article, nous avons proposé une méthode originale de segmentation vidéo interactive. Elle tire son efficacité d'une étape réalisée hors ligne qui produit une sursegmentation en zones quasi-plates de la séquence vidéo et construit un graphe d'adjacence de régions spatio-temporelles. Le processus interactif opère en ligne dans un second temps : l'utilisateur y est impliqué pour guider à l'aide de marqueurs le raffinement de la segmentation et ainsi obtenir les objets d'intérêt.

Nos futurs travaux porteront sur l'optimisation des différentes étapes du processus de segmentation. La parallélisation permettrait de profiter pleinement des processeurs multi-cœurs actuels, notamment pour la partie hors-ligne. En outre, nous souhaitons adapter notre méthode dans un contexte de co-segmentation de séquences vidéo. L'évaluation de notre méthode et sa comparaison à l'état-de-l'art doit aussi être étoffée par l'ajout de critères relatifs à l'évaluation qualitative des résultats finaux de segmentation et à l'évaluation du coût en temps utilisateur nécessaire pour atteindre ces résultats.

Références

[1] R. ADAMS et L. BISCHOF : Seeded region growing. *IEEE Transanction on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[2] F.C. FLORES et R.A LOTUFO : Watershed from propagated markers : An interactive method to morphological object segmentation in image sequences. *Image and Vision Computing*, 28(11):1491–1514, 2010.

[3] K. MCGUINNESS et N.E. O'CONNOR : A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

[4] J.F. RIVEST, S. BEUCHER et J. DELHOMME : Marker-controlled segmentation : an application to electrical borehole imaging. *Journal of Electronic Imaging*, 1(2): 136–142, 1992.

[5] P. SOILLE : Constrained connectivity for hierarchical image partitioning and simplification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(7): 1132–1145, 2008.

[6] J. WEBER, S. LEFÈVRE et P. GANÇARSKI : Video object mining : Issues and perspectives. *In IEEE International Conference on Semantic Computing*, pages 85–90, 2010.

[7] J. WEBER, S. LEFÈVRE et P. GANÇARSKI : Zones quasi-plates spatio-temporelles et segmentation morphologique de séquences vidéo. *In ORASIS - Congrès des jeunes chercheurs en vision par ordinateur*, pages 1–8, 2011.