A propos d'interactions qui permettent d'analyser une vidéo

Axel Carlier, Vincent Charvillat

▶ **To cite this version:**

Axel Carlier, Vincent Charvillat. A propos d'interactions qui permettent d'analyser une vidéo. ORASIS - Congrès des jeunes chercheurs en vision par ordinateur, INRIA Grenoble Rhône-Alpes, Jun 2011, Praz-sur-Arly, France. inria-00596247

HAL Id: inria-00596247
https://inria.hal.science/inria-00596247

Submitted on 26 May 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A propos d'interactions qui permettent d'analyser une vidéo

Axel Carlier Vincent Charvillat

[1] IRIT UMR CNRS 5505 - ENSEEIHT - Université de Toulouse

2 rue Camichel, 31000 Toulouse
{carlier.axel,vcharvillat}@gmail.com

Résumé

*Face à la production massive actuelle d'images et de vidéos numériques, la compréhension automatique de contenus visuels est un enjeu majeur. Quel peut-être le rôle des utilisateurs dans ce processus d'analyse qui s'appuie, aussi, sur des mécanismes de vision artificielle ? Telle est la question centrale soulevée dans cet article. L'approche suivie ici consiste à analyser le comportement des utilisateurs lorsqu'ils visionnent des vidéos interactives (zoomables) pour détecter des régions d'intérêt (ROI). La détection de ROI s'opère ici par analyse des usages (*crowdsourcing*) et non par analyse des propriétés spatio-temporelles du contenu de la vidéo. Notre objectif dans des travaux futurs est de montrer la complémentarité de ces deux méthodes et d'en étudier les modalités.*

Mots Clef

Détection de régions d'intérêt, Vidéos interactives, Apprentissage

Abstract

With a massive amount of new images and videos being produced every day, understanding visual content becomes a major issue in today's research. Can users be part of this process which is usually adressed by computer vision mechanisms ? We try in this paper to answer this question. We present here a method to analyze users' behaviour when viewing interactive videos (zoomable video) to detect regions of interest (ROI). We do not analyze spatio-temporal properties of the video but only traces of users' interactions, to detect the Region of Interest. We aim at investigating in future work the different ways of combining those two methods.

Keywords

Regions of Interest Detection, Interactive video, Machine Learning,

1 Introduction

Pour la plupart des problèmes posés en vision artificielle, une solution entièrement automatique est recherchée. C'est vrai, en particulier, pour les problèmes majeurs que sont la mise en correspondance, la segmentation ou encore la reconstruction tridimensionnelle. Les chercheurs souhaitent, en réponse à ces problèmes, mettre au point des modèles et des algorithmes les plus généraux qui permettent, à partir d'une ou plusieurs images, d'arriver automatiquement à une compréhension correcte de la scène observée. Voilà une gageure face à laquelle les travaux actuels prennent plusieurs directions.

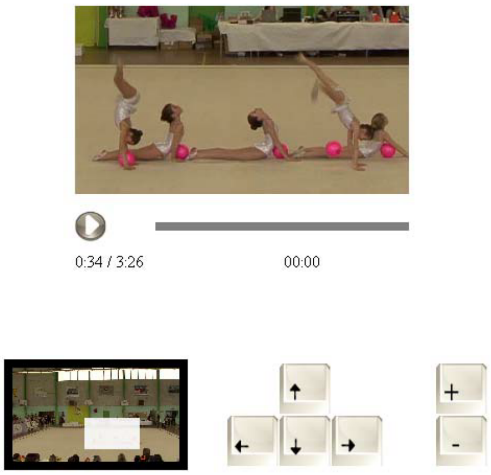

FIGURE 1 – Interface pour l'accès à une vidéo zoomable

Une première attitude est celle où on introduit suffisamment de connaissances sur la scène observée et les capteurs pour se ramener à un problème plus simple pour lequel on arrive à une solution automatique. Ces connaissances forment ce que l'on appelle parfois un *contexte* dans la littérature la plus récente [12]. De la même manière, les méthodes d'*apprentissage artificiel* consistent à se donner une base de connaissances préalables. Cette base d'apprentissage permet à un algorithme de prédire automatiquement des solutions vraisemblables à un problème de reconnaissance d'objets, par exemple. Dans les approches supervisées ou semi-supervisées devenues très populaires dans la

communauté, l'expertise humaine est nécessaire pour établir la base d'apprentissage. C'est une façon élégante et pertinente de mettre au point des algorithmes automatiques après une phase de supervision humaine préalable.

Une seconde direction est celle où, face à des problèmes extrêmement difficiles, on prévoit d'emblée une intervention humaine en vue d'une solution semi-automatique. Pour ce type d'approche, l'enjeu principal est de proposer des interfaces homme-machine bien conçues pour permettre à l'utilisateur d'aider ou de guider efficacement et *explicitement* la résolution d'un problème posé. Nous détaillerons dans la section suivante des travaux sur la segmentation interactive [3, 2] qui illustrent une possible collaboration entre des algorithmes d'analyse du contenu visuel (segmentation via la couleur, la texture) et des mécanismes interactifs (qui permettent d'améliorer la segmentation là où elle est imparfaite).

Dans ce travail, on propose également une approche interactive pour résoudre un problème de détection de régions d'intérêt (ROI) au sein d'une vidéo. Pour cela on choisit d'analyser les comportements des utilisateurs lorsqu'ils visualisent une vidéo pour en déduire automatiquement les ROI [4]. Les utilisateurs de notre système de "vidéo zoomable" utilisent une interface (figure 1) mais ne désignent pas explicitement les ROI. Nous analysons leurs traces de visualisation comme des signaux de retour implicites (*implicit feedback*). En d'autres termes, l'utilisateur "est dans la boucle" d'analyse du contenu visuel mais sans en avoir conscience. Ce travail ouvre des perspectives que nous trouvons intéressantes dans le domaine de l'analyse conjointe du contenu visuel et des usages.

La suite de l'article présente d'abord une analyse bibliographique où nous identifions des relations possibles entre analyse de contenus visuels et analyse des usages interactifs. Dans la seconde section, nous décrivons notre travail récent sur les vidéos interactives visant à déterminer automatiquement les objets visuels d'intérêt présents au sein d'une vidéo. Dans une dernière section, nous proposons de nouvelles pistes de recherche pour combiner plus étroitement encore analyse du contenu visuel et analyse des usages dans le domaine de la vidéo interactive.

2 Etat de l'art

Dans ce paragraphe bibliographique, nous montrons d'abord que l'analyse d'un contenu visuel, comprise au sens de la vision par ordinateur, peut permettre de créer de nouvelles interactions avec une séquence vidéo. Dans un second volet, nous montrons que, réciproquement, l'analyse des interactions vidéos est porteuse de sens et peut permettre d'analyser le contenu visuel. Enfin, combiner ces deux types d'analyse est une piste que certains chercheurs suivent déjà.

2.1 L'analyse de contenu visuel au service de nouvelles interactions

L'article de synthèse de Goldman et al. [8] est une bonne référence qui montre que la détection et le suivi visuel d'objets d'intérêt au sein d'une séquence d'images permettent d'imaginer de nouveaux usages pour la compréhension, l'annotation ou encore l'édition d'une vidéo numérique. Une des fonctionnalités interactives étudiées est appelée "manipulation directe". L'idée est de prendre le contrôle du curseur de temps associé à une séquence d'images en saisissant interactivement un objet visuel d'intérêt qui, au fil de ses déplacement spatiaux dans l'espace image, peut permettre d'avancer ou de reculer temporellement au sein de la vidéo. On parle ainsi d'accès non séquentiel à la vidéo [7]. Il s'agit donc de rendre interactifs des objets visuels associés à des régions d'intérêt (ROI) qu'il faut préalablement détecter. Une ROI est détectable automatiquement dans une vidéo si elle est en mouvement, ou encore parce qu'elle présente un fort contraste colorimétrique avec l'arrière plan : c'est la notion de région saillante [9]. Parmi les modèles d'attention visuelle qui prédisent les régions qui attirent l'oeil, le modèle d'Itti est un des plus cités [10]. Des chercheurs ont également généralisé ce type d'analyse à la dimension temporelle : au sein d'une séquence d'images, des sous-séquences peuvent être identifiées comme plus intéressantes que d'autres. Au-delà de l'application de ces techniques aux résumés automatiques, de nouvelles interactions ont été proposées en tirant parti de cette analyse temporelle du contenu. Le lecteur de vidéo appelé *smart player* [5] propose ainsi d'adapter la vitesse d'affichage d'une vidéo en fonction de l'intérêt du contenu prédit à chaque instant. Les sous-séquences les plus riches et complexes sont ainsi affichées à une vitesse permettant leur interprétation alors que les parties les moins intéressantes défilent plus rapidement.

2.2 Les interactions au service de l'analyse du contenu.

Le travail de Syeda-Mahmood et Ponceleon [18] identifie aussi les sous-séquences les plus intéressantes d'une vidéo mais ces chercheuses opèrent de manière très différente. Elles analysent les interactions des utilisateurs visionnant la vidéo pour en déduire, par apprentissage des comportements les plus fréquents, les sous-séquences d'intérêt. On voit bien ici que l'analyse des usages interactifs permet d'analyser le contenu (approche réciproque à celle du *smart player* [5]) grâce à un grand nombre d'utilisateurs. L'effet d'échelle associé à de tels mécanismes de *crowdsourcing* est une piste crédible pour analyser sémantiquement les contenus visuels [16] ou même reconstruire des scènes [13] grâce à l'*Internet vision*. Les jeux en ligne ou applications web pour l'annotation collaborative d'images ont également montré leur intérêt [20, 21, 15]. Pour le problème de détection de ROI défini ci-dessus, on peut ainsi utiliser l'oculométrie (suivi du regard - *eye-tracking*) pour accumuler des points de fixation d'un grand nombre d'ob-

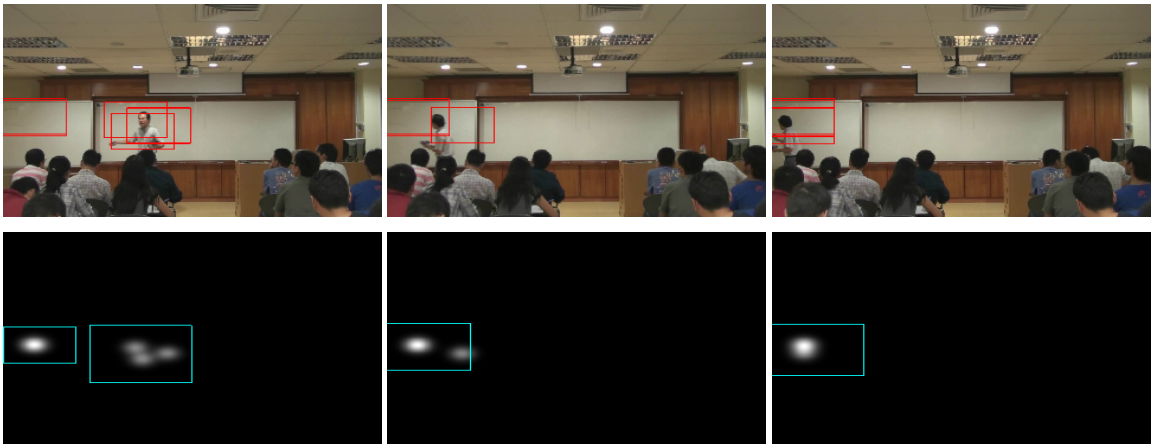

FIGURE 2 – Détections de ROI. Les traces des utilisateurs (première ligne) conduisent aux votes gaussiens accumulés dans une carte de chaleur (seconde ligne) dont l'analyse révèle les ROI (rectangles en cyan).

servateurs et déterminer les objets saillants présents dans une image ou vidéo [19]. Cette approche est néamoins plus intrusive que celles qui consistent à utiliser les traces de visualisation comme des signaux de retour implicites (*implicit feedback*) [18]. On peut observer, dans ce domaine, que plus les interactions sont nombreuses et riches, plus l'exploitation des traces des utilisateurs par *crowdsourcing* est aisée. Dans le travail de Xie et al [22], des images sont visualisées sur des terminaux légers dont les écrans sont de tailles réduites. L'analyse et l'interprétation des agrandissements ou des réductions de la région affichée (par zooms en avant et en arrière), des positionnements de la fenêtre de visualisation (par scrolling) conduisent les auteurs à des cartes d'intérêt. Ces cartes déduites des interactions ressemblent aux cartes de saillance des modèles attentionnels mais sont issues de mesures objectives au même titre que celles provenant de l'oculométrie. Nous avons généralisé ce travail aux vidéos [4]. Nous rappellerons les étapes majeures de ce travail récent dans la suite de cet article. Grâce à une interface permettant de se déplacer et de zoomer dans des vidéos (figure 1), nous pouvons déterminer des ROI par *crowdsourcing*.

2.3 Combinaison d'analyses

Il est naturel de penser à combiner les approches d'analyse du contenu et d'analyse des usages décrites ci-dessus puisqu'elles peuvent fournir des solutions aux mêmes problèmes. La littérature existante [3, 2, 1] ouvre cette voie à travers des travaux récents sur la segmentation interactive d'images. Le modèle de co-segmentation interactive (iCoseg) présenté dans ces travaux consiste à segmenter des objets apparaissant à l'avant-plan d'une scène à partir de plusieurs images distinctes de la scène. L'outil interactif combine de l'analyse de contenu (algorithmes de segmentation automatique par minimisation d'une énergie inspirée des travaux sur l'extraction d'objets visuels à la *object-cutout* [14, 11]) et l'analyse de crayonnages interactifs (*scribbles*) qui permettent de superviser la segmentation. La combinaison se concrétise sous la forme d'un système de recommandation qui suggère à l'utilisateur des zones où la segmentation automatique est imparfaite et où il lui est proposé de superviser le processus. Il nous semble que ce travail montre bien la complémentarité des deux types d'analyses. L'analyse du contenu permet de suggérer, recommander, résumer, filtrer. Les interactions permettent (explicitement ou implicitement) de désigner, corriger, sélectionner. La prise en compte incrémentale ou par renforcement [17] des interactions est un cadre possible pour formaliser cette combinaison d'analyses. C'est avec cet objectif que nous présenterons nos pistes actuelles de recherche (section 4) qui généralisent nos travaux récents permettant de détecter des ROI à partir de l'analyse des usages faits de nos vidéos interactives. Ces travaux sont décrits dans le paragraphe suivant.

3 Détection de ROI via une vidéo interactive

3.1 Détection de ROI

L'objectif de notre travail est la détection automatique de régions d'intérêt, c'est-à-dire des régions spatio-temporelles que les spectateurs sont les plus susceptibles de regarder. On utilise classiquement des modèles attentionnels, comme celui d'Itti ([10]), basés sur des analyses de mouvement ou de composantes lumineuses de l'image (modèle de saillance). Un exemple de l'application de tels modèles est présenté figure 3. L'image supérieure est une carte de saillance, elle détecte les régions présentant un fort contraste colorimétrique avec l'environnement et des régions de couleur "peau" présentant à priori un intérêt. Sur l'exemple de la photo (un gala de gymnastique), de nombreuses régions sont ainsi détectées dont certaines ne sont clairement pas d'intérêt (fenêtres opaques, portes...). L'analyse de mouvement (image inférieure) détecte les

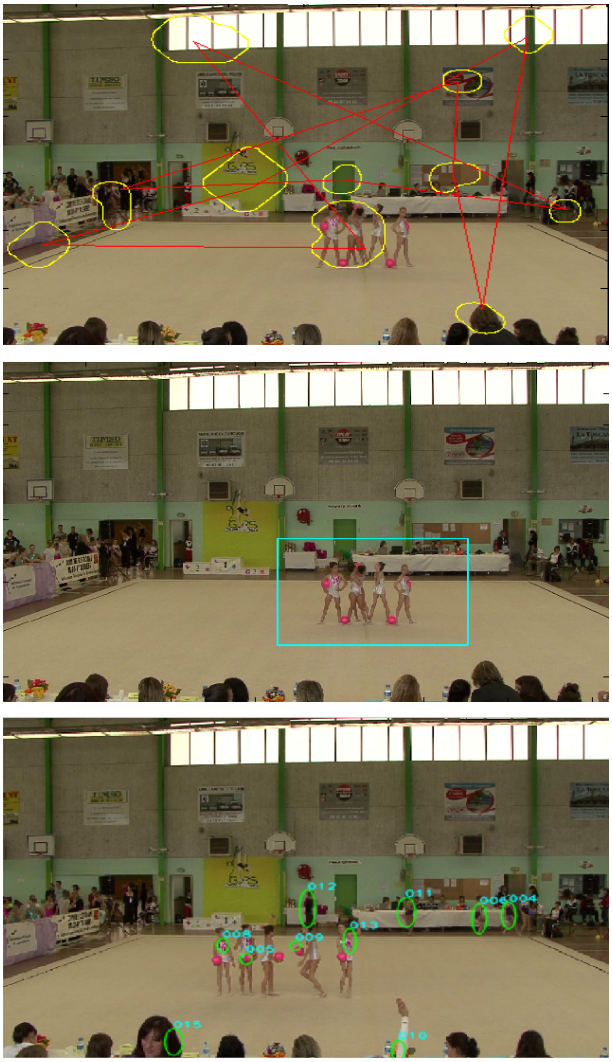

FIGURE 3 – Modèles d'attention visuelle (carte de saillance en haut, carte de mouvements en bas). Notre ROI détectée à partir des traces de visualisation (au centre).

gymnastes, mais aussi les juges au premier plan ou encore des concurrentes se préparant en arrière-plan. Au centre de la figure est illustrée la région détectée par analyse des traces des utilisateurs, qui sont clairement intéressés par les gymnastes seulement. Dans certains cas, les modèles attentionnels semblent donc insuffisants à détecter des ROI. Par définition, les ROI sont des régions qui intéressent les utilisateurs, il est donc naturel d'analyser les interactions de ces utilisateurs avec le contenu visuel afin d'y relever des régions populaires qui seront probablement des ROI. C'est ce que nous avons réalisé dans le travail présenté ci-après.

3.2 Présentation du projet 'video zoomable'

Ce travail trouve son origine dans un projet plus large traitant de la vidéo zoomable. Ce projet était motivé par la problématique suivante : comment streamer des vidéos dont la résolution est si élevée que la bande passante disponible est insuffisante, ou que la taille du dispositif d'affichage est trop limitée ?
Une manière de traiter ce type de problèmes est de mettre en place une architecture permettant de zoomer dynamiquement sur des régions d'intérêt. L'idée est de streamer une version basse résolution du flux vidéo, puis à la demande du spectateur de streamer une sous-région en lui octroyant plus de détails (à une résolution plus élevée) : on obtient ainsi un effet de zoom, d'où le nom de vidéo zoomable. Sélectionner une sous-région d'une vidéo porte le nom anglais de "cropping".
La première partie du projet consistait en l'étude de diverses méthodes de compression vidéo, ayant pour but de limiter l'overhead induit par l'opération de cropping avec des méthodes de compression classiques (de type MPEG4/H.264). La seconde orientation du projet, qui nous intéresse ici, est fondée sur la remarque suivante : les interactions des utilisateurs révèlent les régions ayant un contenu sémantiquement intéressant pour les spectateurs. On peut donc déduire de leurs interactions les régions d'intérêt de la vidéo.

3.3 Interface

Nous avons donc développé une interface pour permettre à des utilisateurs de zoomer dynamiquement sur des vidéos Haute Definition. La figure 1 présente l'interface et ses différentes composantes. En haut on trouve la video centrée sur une région particulière avec un certain niveau de détails choisi par l'utilisateur. Au début de la vidéo, on peut voir toute l'image à basse résolution. L'utilisateur peut alors choisir de zoomer à l'endroit de son choix, soit en pressant le bouton + (le bouton - permet de dézoomer), soit en pointant la souris sur la région qui l'intéresse et en utilisant la molette pour obtenir plus ou moins de détails. A tout moment l'utilisateur est conscient du contexte grâce à la fenêtre située en bas à gauche de l'écran, où la region courante visualisée est représentée par un rectangle blanc sur la frame en entier, affichée en très basse résolution. En cliquant n'importe où sur cette plus petite fenêtre, on centre la région visualisée à l'endroit du clic. Pour modifier la position de la région visualisée, on peut également utiliser les boutons représentant des flèches ou utiliser la souris et faire un drag'n drop directement sur la région visualisée.
Toutes les interactions sont enregistrées dans une base de données de telle sorte que l'on puisse après coup savoir pour chaque utilisateur quelle région a été vue à quel moment.

3.4 Etude d'utilisateurs

Une fois cette interface developpée, nous avons conduit une étude d'utilisation de cette interface. L'idée était d'obtenir des informations sur le comportement des utilisateurs face à une telle interface. Nous avons utilisé un corpus de quatre vidéos, enregistrées à l'aide d'une caméra HD fixe à la résolution 1920x1080 pixels. Trois de ces vidéos montraient un tour de magie, et la quatrième vidéo était extraite d'un gala de gymnastique. En tout, nous avons pu collecter

53 sessions, chaque session correspondant à la vision par l'utilisateur de l'ensemble de la vidéo. Ces 53 sessions ont produit plus de 11 000 interactions.

3.5 Modélisation

La figure 2 montre l'approche générale que nous avons suivie. La ligne supérieure représente quelques images extraites à différents instants d'une vidéo, les rectangles rouges matérialisant les points de vue des spectateurs ayant visionné cette vidéo avec notre interface. La première étape consiste en la création d'une carte de chaleur pour chaque image (heatmap), visibles sur la deuxième ligne de la figure 2. Les régions claires de ces heatmaps sont des points chauds (hotspots) sur lesquels plusieurs utilisateurs se sont concentrés. L'analyse de ces heatmaps nous permet de déterminer les regions d'intérêt (rectangles bleus) implicitement indiquées par les utilisateurs.
On commence donc par créer des cartes de chaleur en cumulant des votes gaussiens correspondant aux points de vue des utilisateurs. L'utilisation de gaussiennes nous a paru appropriée dans ce contexte car la modélisation des cartes de chaleur s'en trouve par la suite facilitée.
La figure 5 suivante montre l'empreinte gaussienne des votes centrés sur une fenêtre de visualisation. Pour chaque image de la séquence considérée, on cumule alors les votes issus des points de vue adoptés par les utilisateurs. On obtient une carte de chaleur. Pour modéliser cette carte de chaleur, un mélange de gaussiennes est légitime selon une densité de forme générale :

$$p(\mathbf{x}|\theta_t) = \sum_{j=1}^{K} \omega_{t,j} \;\; p(\mathbf{x}|\mu_{t,j}, \Sigma_{t,j}) \qquad (1)$$

où $\omega_{t,j}$ sont les poids relatifs des K ROIs considérées à l'image t avec $\omega_{t,j} > 0$ et $\sum_{j=1}^{K} \omega_{t,j} = 1$. De même $\mu_{t,j}$ et $\Sigma_{t,j}$ sont les moments de la j^{ieme} gaussienne à l'instant t et ces paramètres sont regroupés dans un vecteur $\theta_t = \{\omega_{t,1} \ldots \omega_{t,K}, \mu_{t,1} \ldots \mu_{t,K}, \Sigma_{t,1} \ldots \Sigma_{t,K}\}$.
Pour estimer ces paramètres θ_t *de manière robuste*, nous commençons par appliquer un algoritme de clustering, l'algorithme Mean-Shift. On commence par générer un ensemble de points suivant la distribution décrite par la carte de chaleur, puis l'on regroupe ces points en différents clusters grâce à Mean-Shift [6]. L'intérêt de cet algorithme est qu'il ne présuppose pas de connaissance préalable du nombre de clusters, et donc du nombre de gaussiennes [1].
Un exemple d'une exécution de Mean-Shift peut être vu en bas de la figure 6. Dans cet exemple, l'algorithme détecte deux clusters différents, et en fournit les modes (associés aux moments $\mu_{t,j}$), ainsi que les poids relatifs (calculés simplement au moyen du nombre de points d'un cluster divisé par le nombre de points total).
Il reste à déterminer la matrice de variance-covariance de chaque cluster, ce qui pourrait être fait de manière classique en utilisant un estimateur au maximum de vraisemblance. Dans notre travail nous avons décidé d'utiliser un estimateur robuste MCD (Minimum Covariance Determinant) qui nous semblait plus précis, car s'affranchissant des éventuelles données aberrantes (par exemple des votes marginaux). Au final ce choix nous permet de récupérer des régions d'intérêt plus localisées, ce qui dans notre application était un avantage.
A l'issue de cette phase d'estimation, nous utilisons les paramètres des gaussiennes estimées pour reconstruire des points de vue correspondant, qui forment finalement les ROI détectées (haut de la figure 6).

1. Seule une largeur de bande permettant de fusionner deux ROI est à fournir.

3.6 Résultats et applications de la détection de ROI

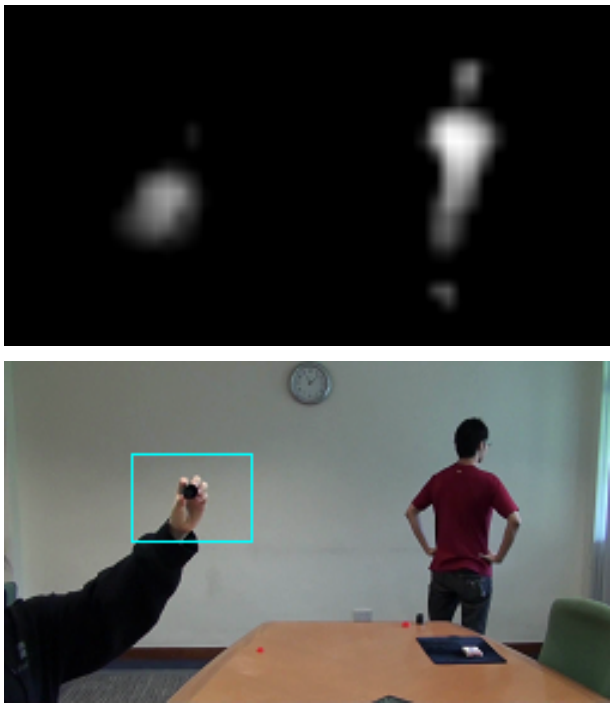

FIGURE 4 – Carte de saillance (en haut) et détection de ROI par notre méthode (en bas) sur une image extraite d'un tour de magie.

La figure 4 ci-dessus illustre un de nos résultats forts : les interactions des utilisateurs révèlent parfois des régions indétectables par les méthodes classiques comme la carte de saillance. Dans cet exemple d'une vidéo montrant un tour de magie, l'image présentée arrive après une phrase du magicien suggérant au spectateur de bien regarder la valeur du dé montré sur la partie gauche de l'écran. C'est donc assez naturellement que la ROI détectée par notre méthode est localisée en ce dé (le rectangle bleu sur l'image du bas de la figure 4). Dans ce cas précis, l'analyse de contenu (par le biais d'une carte de saillance) détecte des régions comme le T-Shirt du personnage de droite ou le bras du personnage de gauche, mais pas le dé (image du haut de la figure 4). Ce

résultat montre bien que l'analyse des interactions des utilisateurs peut fournir un résultat sensiblement meilleur que l'analyse de contenu.
Dans le travail présenté précédemment, la détection automatique n'était qu'une étape vers la génération d'une nouvelle vidéo de plus basse résolution que l'originale, mais conservant les détails essentiels à la compréhension du contenu. En effet dans le cas de vidéos Haute Définition, on se heurte à des problèmes de taille d'affichage (les Smartphone par exemple ont une résolution limitée) mais également à des problèmes de bande passante. Pour profiter malgré tout du contenu sans forcément perdre les détails gagnés grâce à la Haute Définition, un processus de Video Retargeting peut être utile. la détection de ROI intervient lorsque, à certains moments dans la vidéo, l'utilisateur préfère voir une région plus localisée de la vidéo avec plus de détails, plutôt que tout le contenu à basse résolution (auquel cas il pourrait ne pas distinguer certains détails importants). La détection de ROI a également des applications dans la compression vidéo, par exemple dans les algorithmes de compression par ROI. ces algorithmes implémentent des stratégies d'encodage plus fin des régions d'intérêts, et plus grossier des régions de non-intérêt comme l'arrière-plan. La détection automatique des ROI est un enjeu crucial de ce type d'algorithme, et toute contribution permettant de détecter plus efficacement les ROI peut améliorer la qualité de la compression.

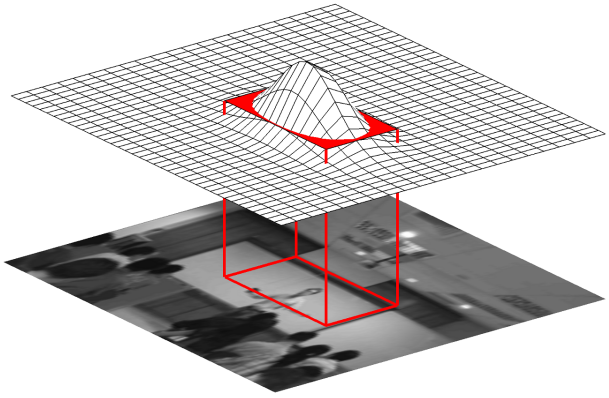

FIGURE 5 – Vote selon une empreinte gaussienne.

4 Travaux futurs

Il nous semble intéressant de capitaliser sur le travail réalisé afin de proposer une nouvelle étude des possibles complémentarités entre l'analyse du contenu et l'analyse par l'usage.
Une première piste est à explorer du côté des nouvelles interactions rendues possibles par une analyse de contenu préalable sur la vidéo regardée par un utilisateur. Une telle analyse de contenu pourrait révéler des régions d'intérêt qu'il serait alors possible de "proposer" aux utilisateurs. De même, un problème récurrent de notre interface était

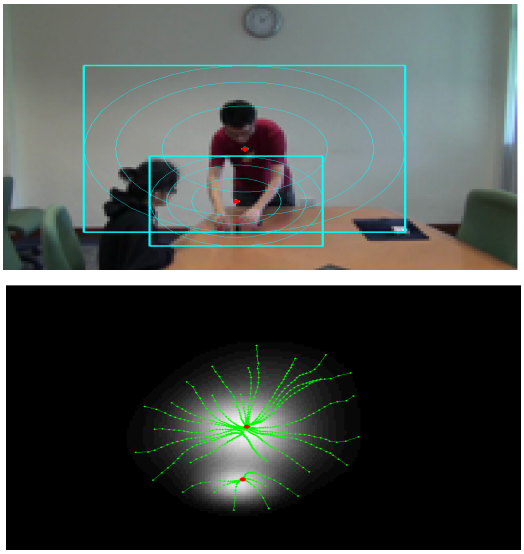

FIGURE 6 – Modes détectés par Mean Shift et détection de ROI (haut) après de multiples démarrages (bas).

la perte régulière de l'objet d'intérêt. Il est en effet assez peu naturel de suivre un objet en mouvement avec notre interface. On pourrait imaginer être capable de suivre automatiquement le mouvement d'un objet si celui-ci a été détecté lors de l'analyse de contenu.
Ces interactions sont matérialisées sur la figure 7. La première ligne reprend une interface similaire à celle dont nous avons déjà parlé, mais y inclut une nouvelle interaction rendue possible par l'analyse de contenu. Lorsque l'utilisateur survole la vidéo avec sa souris, il voit apparaître un cadre noir si la souris se situe sur une région d'intérêt. Il lui suffit alors de cliquer à l'intérieur de ce cadre pour voir la région zoomée. L'avantage de cette méthode est que l'analyse de contenu peut nous permettre de suivre automatiquement la région en cas de mouvement de l'objet, ce qui élimine le défaut de la perte de l'objet d'intérêt décrit ci-dessus.
Dans la deuxième ligne de la figure, l'interface adoptée est complètement différente. A tout moment, l'utilisateur visualise les régions d'intérêt de la vidéo sur une barre située dans la partie inférieure de l'interface. En survolant ces régions avec la souris, l'utilisateur les voit se matérialiser par un rectangle noir (deuxième image), et il lui suffit de cliquer pour zoomer sur la région désirée. On remarque que dans ce cas, la frame toute entière vient remplacer la région zoomée dans la barre de la partie inférieure, ce qui permet à l'utilisateur de revenir d'un seul clic à la vue globale.
Notre objectif est également de renforcer l'analyse de contenu en lui adjoignant des données récupérées des interactions des utilisateurs. On peut par exemple sélectionner parmi les régions d'intérêt détectées lors de l'analyse de contenu seulement celles qui ont été validées par les utilisateurs, c'est-à-dire celles qui ont effectivement été visionnées. Si des régions apparaissent intéressantes aux yeux

d'assez d'utilisateurs, on peut également les ajouter aux régions proposées (dans le cas où l'analyse de contenu ne les avait pas détectées).

Nous prévoyons dans un premier temps de retravailler notre interface de vidéo zoomable pour l'améliorer en utilisant l'analyse de contenu, puis de réaliser une étude d'utilisateurs. A partir des résultats de cette première étude, nous espérons pouvoir combiner efficacement analyse de contenu et traces utilisateurs pour parvenir à une deuxième interface, où la quantité d'interaction demandée aux utilisateurs sera moindre. Une deuxième étude d'utilisateurs viendra confirmer ou infirmer ce résultat. Nous espérons pouvoir intégrer de premiers résultats dans la version finale éventuelle de cet article.

Au final nous espérons obtenir des résultats probants. Une première satisfaction serait d'observer l'utilisation des nouvelles interactions permises par l'analyse de contenu. Ce serait une preuve de plus de l'utilité de l'analyse de contenu dans l'amélioration des interfaces homme-machine.

5 Conclusion

En guise de conclusion, disons d'abord, tout aussi honnêtement que simplement, que cet article n'est pas un article de vision par ordinateur. Pourtant, il nous semble important de soumettre ces idées situées à la frontière de notre discipline car, indiscutablement, les outils de vision artificielle sont et seront intégrés à de nombreux systèmes interactifs et autres applications multimodales. Les références bibliographiques que nous avons citées sont issues de nombreuses communautés mais elles ont toutes en commun de placer les utilisateurs au centre de la réflexion. Dans le cadre de l'interprétation visuelle globale de scènes variées, il nous semble que les utilisateurs auront nécessairement un rôle à jouer en amont ou en aval des composants de vision qui deviennent matures dans les domaines de la reconstruction géométrique ou encore de la reconnaissance d'objets.

Remerciements

Nous remercions les co-auteurs de l'article présentant l'intégralité de nos recherches sur les vidéos zoomables [4], les étudiants qui contribuent au développement des nouvelles interfaces vidéos et nos collègues singapouriens de NUS sans qui ce travail n'aurait pas pu commencer.

Références

[1] W.-C. C. T. C. Adarsh Kowdle, Dhruv Batra. imodel : Interactive co-segmentation for object of interest 3d modeling. In *Workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments, European Conference on Computer Vision*, 2010.

[2] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg : Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176, 2010.

[3] D. Batra, D. Parikh, A. Kowdle, T. Chen, and J. Luo. Seed image selection in interactive cosegmentation. In *ICIP*, pages 2393–2396, 2009.

[4] A. Carlier, V. Charvillat, W. T. Ooi, R. Grigoras, and G. Morin. Crowdsourced automatic zoom and scroll for video retargeting. In *ACM Multimedia*, pages 201–210, 2010.

[5] K.-Y. Cheng, S.-J. Luo, B.-Y. Chen, and H.-H. Chu. Smartplayer : User-centric video fast-forwarding. In *ACM CHI 2009 Conference Proceedings*, 2009.

[6] D. Comaniciu and P. Meer. Mean shift : A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5) :603–619, 2002.

[7] P. Dragicevic, G. Ramos, J. Bibliowitcz, D. Nowrouzezahrai, R. Balakrishnan, and K. Singh. Video browsing by direct manipulation. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 237–246, New York, NY, USA, 2008. ACM.

[8] D. B. Goldman, C. Gonterman, B. Curless, D. Salesin, and S. M. Seitz. Video object annotation, navigation, and composition. In *UIST*, pages 3–12, 2008.

[9] J. Han, K. N. Ngan, M. Li, and H. Zhang. Unsupervised extraction of visual attention objects in color images. *IEEE Trans. Circuits Syst. Video Techn.*, 16(1) :141–145, 2006.

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11) :1254–1259, 1998.

[11] Y. Li, J. Sun, and H.-Y. Shum. Video object cut and paste. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 595–600, New York, NY, USA, 2005. ACM.

[12] O. Marques, E. Barenholtz, and V. Charvillat. Context modeling in computer vision : techniques, implications, and applications. *Multimedia Tools and Applications*, 51 :303–339, 2011. 10.1007/s11042-010-0631-y.

[13] M. G. R. S. Noah Snavely, Ian Simon and S. M. Seitz. Scene reconstruction and visualization from community photo collections. *Proceedings of the IEEE*, 98(8) :1370–1390, 2010.

[14] C. Rother, V. Kolmogorov, and A. Blake. "grabcut" : interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3) :309–314, 2004.

[15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme : A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3) :157–173, 2008.

[16] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu. Watch what i watch : using community activity to understand content. In *Multimedia Information Retrieval*, pages 275–284, 2007.

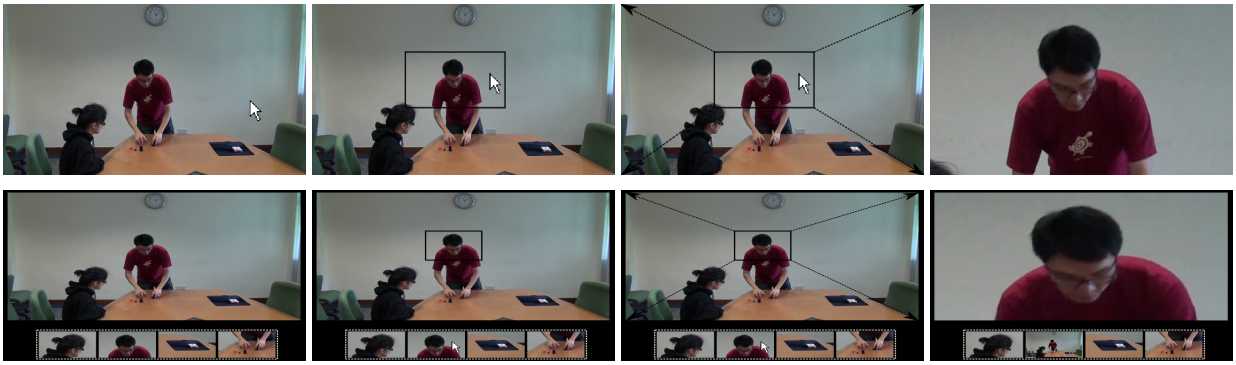

FIGURE 7 – Deux nouveaux types d'interface permettant de combiner analyse de contenu et analyse des usages. Une ROI, détectée par analyse de contenu, est rendue interactive (première ligne). Les imagettes présentant des ROI candidates sont présentées simultanément et sont également rendues cliquables (seconde ligne).

[17] R. S. Sutton and A. G. Barto. Reinforcement learning : An introduction. *IEEE Transactions on Neural Networks*, 9(5) :1054–1054, 1998.

[18] T. Syeda-Mahmood and D. Ponceleon. Learning video browsing behavior and its application in the generation of video previews. In *Proceedings of MULTIMEDIA '01*, pages 119–128, Ottawa, Canada, 2001. ACM.

[19] N. Ukita, T. Ono, and M. Kidode. Region extraction of a gaze object using the gaze point and view image sequences. In *ICMI '05 : Proceedings of the 7th international conference on Multimodal interfaces*, pages 129–136, Torento, Italy, 2005. ACM.

[20] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI*, pages 319–326, 2004.

[21] L. von Ahn, R. Liu, and M. Blum. Peekaboom : a game for locating objects in images. In *CHI*, pages 55–64, 2006.

[22] X. Xie, H. Liu, S. Goumaz, and W.-Y. Ma. Learning user interest for image browsing on small-form-factor devices. In *Proceedings of CHI '05*, pages 671–680, Portland, Oregon, USA, 2005. ACM.