Résumé automatique de contenus multimédia

Etudiants : Maxime Bichon, Maxime Rivière, Arthur Crenn
Encadrant pédagogique : Ewa Kijak
Encadrant Entreprise : Claire-Hélène Demarty

***Résumé*—À l'initiative de Technicolor, entreprise leader dans les domaines de la vidéo et de l'imagerie numérique, ce projet a pour objectif la réalisation d'un logiciel résumant automatiquement un ensemble de vidéos. La récente multiplication des moyens d'acquisitions pour la vidéo (smartphones, caméras, appareils photos, ...), a entraîné une croissance de la quantité de données que les particuliers ont à traiter. La principale mission du logiciel développé est de déterminer des caractéristiques intéressantes des vidéos et des images, en déduire les moments intéressants du contenu multimédia et enfin de construire à partir de ces moments un rendu innovant et artistique.**

***Mots clés*—Résumé, Vidéos, Classification, Caractéristiques, Rendu, multimédia, automatique, mosaïque, clustering.**

I. Introduction

A. Contexte

Lors d'un événement, que ce soit un concert (cf. Figure 1), une rencontre sportive ou un mariage, il est courant d'obtenir de nombreuses données et de souhaiter les synthétiser. Cette synthèse se fait généralement via un montage impliquant un tri des moments intéressants ainsi qu'une mise en forme plus ou moins artistique des passages conservés. Les problèmes majeurs sont une étape de tri subjective, longue et rébarbative ainsi qu'un montage exigeant des compétences spécifiques pour utiliser des logiciels, parfois difficiles à prendre en main.

FIGURE 1. Exemple de contenus acquis lors d'un concert

B. Objectifs

L'objectif de ce projet est de fournir aux particuliers un moyen de réaliser les deux étapes de tri et mise en forme d'un montage vidéo d'une manière cohérente, artistique et automatique. C'est en s'aidant des connaissances déjà acquises dans le domaine de l'imagerie numérique et des travaux déjà effectués par la communauté scientifique que nous cherchons un moyen de répondre aux différentes problématiques soulevées par le sujet. Dans un premier temps nous présenterons les caractéristiques que nous extrayons des vidéos ou des images. A partir de ces éléments nous avons déterminé lesquels sont jugés importants et comment ils peuvent être utilisés afin de classer les différents moments en fonction de leur importance. Pour finir, nous avons recherché une méthode innovante permettant de présenter ces différents moments sous forme de résumé.

C. État de l'Art

Le résumé automatique de contenus multimédia est un sujet de recherche qui se développe, et plusieurs méthodes[1][2][3][4] ont été proposées afin de définir et d'identifier le contenu le plus important d'une vidéo. La plupart de ces approches permettent d'aboutir à des résumés optimaux d'un point de vue mathématique. En effet, les résumés produits le sont en cherchant à respecter des critères scientifiques qui ne sont pas forcément représentatifs de critères plus perceptuels et difficiles a définir comme : un rendu artistique, un rendu qui plait. Actuellement la recherche dans ce domaine ne sait pas encore parfaitement définir concrètement et quantitativement des notions très subjectives et perceptuelles, qui pourtant devraient être les critères choisis pour la phase d'évaluation. De plus, une autre problématique de ce domaine de recherche est de juger la qualité du résumé généré. Ainsi, il n'existe pas encore de mesures objectives permettant d'estimer l'impact des critères choisis dans la sélection des moments intéressants. De plus, les ébauches de métrique de qualité suffisamment proche du système visuel humain ne sont pas encore abouties afin d'assurer un rendu agréable et efficace. C'est pourquoi les tests utilisateurs sont une alternatives à l'évaluation par une métrique.

II. Description du projet

Le cycle de traitement effectué par le logiciel est décomposé en trois étapes : Extraction des caractéristiques, Agrégation et Rendu. Le schéma ci-dessous décrit la chaîne de traitement.

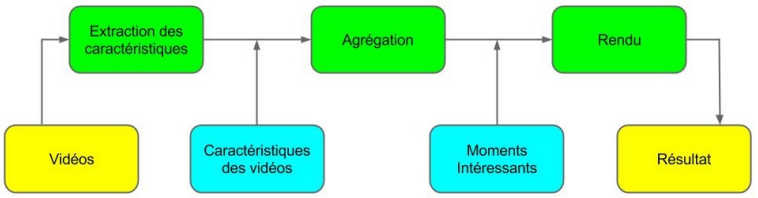

FIGURE 2. Schéma résumant l'architecture de notre projet

A. Extraction des caractéristiques

Cette première partie sert de base au traitement effectué par le logiciel. C'est à partir de l'ensemble des vidéos que souhaite résumer l'utilisateur que nous extrayons les informations utiles à la détection de moments intéressants, réalisée dans la seconde partie. Dû à l'importance en terme de charge de travail du projet, un certain nombre de caractéristiques ont été extraites sous différents formats (Texte, XML, ...) par Technicolor. Une première étape a donc été de sélectionner les caractéristiques les plus utiles et d'en effectuer l'analyse afin de les utiliser au sein de notre programme. A l'issue de cette étape, nous avons décidé d'utiliser les données suivantes :

1) Activité : Cette caractéristique utilise une analyse des mouvements présents et de leur amplitude d'une image à l'autre afin de quantifier la vivacité d'une action présente dans la séquence étudiée.

2) Détection de Visages : Cette donnée nous permet de savoir si des visages sont présents dans une image ainsi que leur nombre. Une question importante est de savoir si déterminer la simple présence d'un ou plusieurs visages est suffisant (utilisé comme une valeur binaire), ou si le nombre de visages a une importance non négligeable.

3) Esthétique : Cette caractéristique est une concaténation des données bas niveau extractibles d'une image permettant d'obtenir un critère objectif de la "beauté" d'une image. Une description de cette caractéristique est fournie de façon plus détaillée dans[5].

4) Saillance : Cette information détecte l'importance d'éléments internes à l'image qui prennent l'ascendant sur les autres éléments. En d'autres termes elle définit la présence de données qui captent l'attention. Un cas concret serait par exemple l'utilisation d'un texte fortement coloré, plaqué sur une image. Le texte serait un élément saillant car il attire immédiatement le regard.

B. Agrégation des caractéristiques

Cette partie du programme utilise les données issues de l'extraction précédemment expliquée. Elle se décompose en deux étapes : l'évuluation des caractéristiques et la sélection des moments intéressants. Pour les quatre informations décrites ci-dessus, nous obtenons des valeurs numériques uniques par image. L'évaluation consiste à analyser ces valeurs dans le but d'attribuer un score d'importance à chaque image. Puis, nos vidéos étant segmentées en plans, nous cherchons à agréger ces scores par images pour obtenir un seul score par plan. La sélection pour le rendu final se fera en effet parmi les différents plans des vidéos. Nous présentons ci-dessous les différentes versions d'évaluation que nous avons utilisées.

- Première méthode d'évaluation
  Tout d'abord, nous avons choisi de normaliser les valeurs associées à une vidéo l pour une caractéristique donnée k et ceci pour toutes les images i. Cette normalisation permet d'avoir des données comparables. La formule ci-dessous permet de normaliser les valeurs des caractéristiques où v représente la valeur d'une caractéristique pour une image et Vn est la valeur normalisée de v.

$$Vn_{(i,k,l)} = \frac{v_{(i,k,l)} - min_{k,l}}{max_{k,l} - min_{k,l}} \tag{1}$$

  Ainsi, dans la première version de notre normalisation chaque donnée est traitée de manière indépendante. Une fois les valeurs normalisées obtenues à partir de nos caractéristiques pour chaque image, nous avons décidé d'attribuer un score à chaque plan j via la formule suivante où S correspond à la note d'un plan :

$$S_k^l(j) = \frac{1}{N}\sum_{i=0}^{N} Vn_{(i,k,l)} \tag{2}$$

  où N est le nombre d'images de la séquence courante.
  A ce stade, nous avons accès au score d'un plan en fonction d'une certaine caractéristique. Il ne nous reste plus qu'à évaluer chaque séquence en prenant en compte toutes nos caractéristiques. Pour se faire, nous avons décidé d'additionner le score de chaque information associée à un plan ce qui est donné via la formule ci-dessous :

$$S_{(j,l)} = \sum_{k=0}^{K} S_{(j,k,l)} \tag{3}$$

- Seconde méthode d'évaluation
  La différence majeure de cette méthode comparée à la première version est que nous allons traiter chaque information de manière globale. C'est-à-dire qu'au lieu de normaliser les valeurs de nos caractéristiques par image et par vidéo, nous allons maintenant normaliser chacune de ces valeurs en prenant en compte toutes les vidéos. Ainsi, le score de chaque image sera normalisé en fonction de l'ensemble des images traitées. L'équation (1) devient donc :

$$Vn_{(i,k)} = \frac{v_{(i,k)} - min_k}{max_k - min_k} \tag{4}$$

- Troisième méthode d'évaluation

  Dans cette version, nous allons évaluer les caractéristiques par rapport à leur moyenne obtenue sur toutes les vidéos.

$$moy_k = \frac{1}{N}\sum_{l=0}^{L}\sum_{i=0}^{N_l} Vn_{(i,k)} \tag{5}$$

  Cette moyenne va maintenant, nous servir à calculer la distance entre celle-ci et la valeur d'une image pour une caractéristique. Ainsi, on va accorder de l'importance à ce qui est rare. En effet, plus la distance va être élevée plus l'image aura une valeur plus importante. Voici la formule utilisée :

$$Vn_{i,k} = \mid Vn_{i,k} - moy_k \mid \tag{6}$$

Pour finir, notre méthode de sélection a toujours été la même. C'est-à-dire que nous avons décidé de trier les valeurs

des scores de chaque plan de toutes les vidéos d'entrées par ordre décroissant. Ensuite, nous prenons les N premières séquences de cette liste triée afin d'obtenir une vidéo de deux minutes. Nous avons jugé pertinent qu'un résumé dépassant cette durée devenait trop long et Technicolor nous a conforté dans ce choix.

C. *Clustering de plans*

Dans le but d'éviter une répétition des plans semblables et afin de faciliter l'implémentation d'un possible apprentissage pour remplacer la phase de sélection par calcul d'un score global, nous avons mis en place un clustering basé sur des caractéristiques visuelles des images. Elle se déroule de la manière suivante : tout d'abord, nous allons parcourir chaque plan de toutes les vidéos. Pour le plan courant, p, nous allons déterminer son image représentative, I_p. Nous avons choisi de sélectionner l'image du plan p possédant le plus fort score. La seconde étape de cet algorithme consiste à calculer un vecteur représentant le descripteur visuel de I_p. Nous avons le choix entre deux méthodes :

- Calculer un nouveau descripteur visuel comme l'histogramme d'orientations du gradient ou l'histogramme des couleurs.
- Réutiliser comme descripteur l'une des caractéristiques précédentes déjà extraites.

Une fois le descripteur, $D(I_p)$, calculé, nous allons maintenant appliquer l'algorithme du K-means[7] sur chaque $D(I_p)$ de chaque plan.

A l'issue de cette étape, l'agglomération visuelle a produit une classification des différentes séquences présentes dans les données.

Il nous reste à choisir pour chaque classe son meilleur représentant. Deux options sont possibles pour faire cela :

- Sélectionner l'image I_p la plus proche du barycentre de chaque groupe.
- Sélectionner l'image I_p du groupe obtenant le meilleur score d'importance lors de l'évaluation.

Pour finir, la dernière étape est de générer le résumé final. Ainsi, nous faisons un résumé des groupes en utilisant uniquement des séquences p dont l'image I_p est représentative d'un groupe. L'objectif de cette étape est d'obtenir un panel varié de séquences dans le résultat final. Ainsi des plans jugés peu intéressants par l'évaluation seront peut-être présents, mais on obtiendra un rendu hétérogène en incorporant des séquences diverses. A l'heure de l'écriture de cet article, nous n'avons pas encore pu tester cette méthode.

D. *Rendu*

Cette partie est l'aboutissement des phases précédentes, elle va permettre de concaténer les différentes séquences vidéo sélectionnées pendant l'étape d'agrégation. Le rendu peut être de différente forme mais dans tous les cas il ne doit pas dépasser deux minutes, sous format vidéo. Nous avons été très libres dans cette phase du projet puisque l'on nous a demandé d'être le plus innovant possible. Il s'agit donc ici plus d'expérimentation que de mise en pratique de techniques de rendu traditionnel.

Le rendu n'étant pas obligatoirement partie intégrante du logiciel à rendre, nous avions l'autorisation d'utiliser des logiciels extérieurs afin de donner des méthodes différentes.

Voici les différents méthodes auxquelles nous avons pensées :

- Stylisation de vidéos : il s'agit ici de transformer le résumé de vidéos réelles en une forme de dessin non-réaliste, cf. Figure 3.

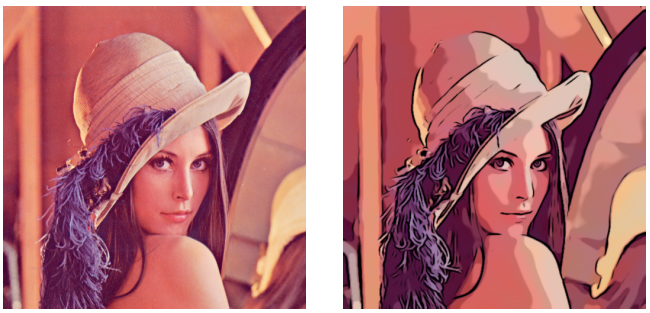

FIGURE 3. Image originale de Lena à gauche et image stylisée à droite

- Folioscope : l'objectif de cette technique est de prendre directement les images les plus importantes et de les faire défiler à la manière d'un folioscope. La vitesse de défilement serait plus faible, puisque les images ne sont pas forcement corrélées entre elles. Une technique similaire est utilisée dans les films de la société Marvel, avec un défilement rapide des pages d'un comic book.

FIGURE 4. Exemple de folioscope

- Motion Magnification : cela correspond à l'utilisation d'une vidéo sous différents angles très similaires afin de simuler un effet de 3D dans la vidéo[8][9].
- Système de particules intéractif : dans le cas où l'utilisateur pourrait interagir avec l'application, le lancé de particules permettrait à celui-ci de choisir quelle vidéo il veut regarder en priorité. Chaque particule correspond à une séquence vidéo du résumé. En voyant une particule qui l'intéresse, l'utilisateur pourrait cliquer dessus afin de visualiser la séquence entièrement avant de passer ensuite à une autre vidéo.
- Vieux cinéma : l'objectif de cette technique est de donner un effet vieux film au résumé obtenu. Les techniques envisagées afin d'obtenir ce résultat consistent principalement à l'ajout de deux bandes noires dans les parties supérieure et inférieure de la vidéo, ainsi qu'une augmentation de la saturation des images.
- Diaporama : celui-ci est un simple défilement des meilleurs images, selon les scores attribués lors de l'évaluation. La vitesse de défilement est relativement

lente, permettant au spectateur d'analyser en profondeur le contenu.
- Mosaïque d'images : il s'agit simplement d'une mosaïque d'images. Celles-ci suivent le même principe que le diaporama.
- Cubico-Magico : c'est l'amélioration de la mosaïque d'image en 3D. Chaque image est remplacée par un cube dont les faces sont une image. Il y a donc une gamme plus importante d'images qui pourraient chacune traiter d'un thème plus précis. Le cube tourne sur lui-même au bout d'un certain temps afin de passer au thème suivant.
- Rouleau d'images : c'est un cylindre vertical divisé en plusieurs parties, chacune correspondant à une séquence du résumé. Chaque fenêtre ainsi présentée sur le cylindre contient la première image de la séquence. Le cylindre tourne sur lui-même et à chaque nouvelle image, celle-ci s'agrandit afin de montrer la séquence vidéo en plein écran.
- Réalité augmentée : il serait possible de voir le résumé sous forme 3D par l'intermédiaire de lunettes 3D ou d'un appareil permettant de visualiser la scène.

III. Résultat

Notre base de données est composée de trois événements. Un concert de Muse que Technicolor nous a fourni, le Challenge Armoricain Supélec ou C.A.S. (compétition sportive réunissant plusieurs écoles d'ingénieurs du Grand Ouest) et la Remise Des Diplômes de l'ESIR de l'année 2014-2015 (R.D.D).

	Concert de Muse	**R.D.D.**	**C.A.S.**
Nombre de vidéos	60	11	14
Durée totale	5h00m53s	1h38m41s	22m05s
Taille totale	2.26 Go	9.79 Go	1.51 Go
Nombre de frames	415 236	58 024	40 318
Nombre de shots	7 043	134	483

TABLE I
PRÉSENTATION DE NOTRE BASE DE DONNÉES

Avant d'enchaîner sur la présentation de nos résultats, il est nécessaire de montrer un exemple de ce que l'on appelle une mauvaise image en terme de score, de même pour une bonne image.

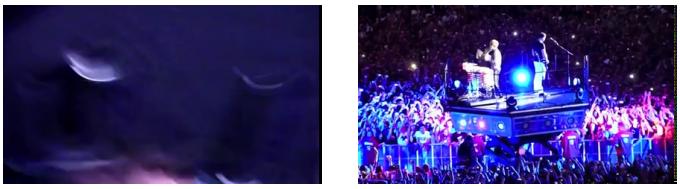

FIGURE 5. Image avec un mauvais score à gauche et image avec un bon score à droite

Ainsi, on peut voir que les images floues ne seront pas prises dans notre résumé ce qui nous permet d'obtenir une vidéo plutôt regardable.

La figure 6 présente la mosaïque d'images obtenue en prenant les N meilleures images de notre résumé automatique :

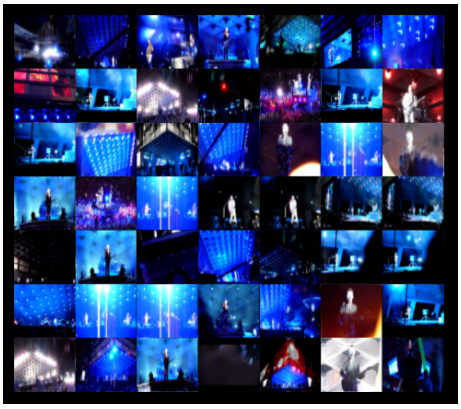

FIGURE 6. Exemple de mosaïque sur le concert de Muse où N = 49

Au moment où cet article est écrit, nous n'avons pas encore implémenté le clustering qui permettra d'obtenir une mosaïque beaucoup plus hétérogène. En effet, comme on peut voir ci-dessus, notre mosaïque possède plusieurs images issues d'un même plan Un autre type de rendu que nous avons implémenté est le Cubico-Magico qui nous permet d'obtenir plusieurs mosaïques d'images en 3D. En combinant ce style de rendu avec le clustering, nous allons obtenir un cube dont chaque face traitera d'un thème. Ainsi, on peut voir que notre clustering de plans nous permettra d'améliorer nos méthodes de rendus en plus de nous rendre indépendant des caractéristiques fournies par Technicolor.

IV. Conclusion

Tout au long de ce papier nous avons pu voir que le sujet est complexe et demande beaucoup de travail. Néanmoins nos premiers résultats sont probants et tous les acteurs sont satisfaits du déroulement de ce projet. Comme énoncé au début de ce papier, l'une des problématiques de ce domaine de recherche est de juger la qualité du résumé généré. En effet, nous arrivons à obtenir différents types de résumés en fonction de la méthode utilisée cependant nous n'avons pas eu le temps de mettre en place une routine de tests utilisateurs afin d'avoir une évaluation plus subjective. Au niveau des méthodes de rendus, nous avons pu en tester quelques-unes, les méthodes restantes serviront peut-être ultérieurement à l'entreprise. Pour finir, nous sommes ravis d'avoir pu travailler avec Technicolor sur un sujet aussi complexe et innovant que le résumé automatique de contenus multimédias bien que nous n'ayons pas pu implémenter toutes nos idées.

Références

[1] Bernard Merialdo Itheri Yagianoui and Benoit Huet, "Automatic video summarization," .

[2] Itheri Yagianoui Bernard Merialdo, Benoit Huet and Fabrice Souvannavong, "Automatic video summarization," .

[3] Courtenay Cotton Wei Jiang and Alexander C. Loui, "Automatic consumer video summarization by audio and visual analysis," .

[4] Chih-Jen Lin Aditya Khosla, Raffay Hamid and Neel Sundaresan, "Large-scale video summarization using web-images priors," .

[5] Jia Li Ritendra Datta, Dhiraj Joshi and James Z. Wang, "Studying aesthetics in photographics images using a computational approach," .

[6] Chiu-Sing CHOY Wei HAN, Cheong-Fat CHAN and Kong-Pang PUN, "An efficient mfcc extraction method in speech recognition," .

[7] Tapas Kanungo et al., "An efficient k-means clustering algorithm : Analysis and implementation," .

[8] Hao-Yu Wu et al., "Eulerian video magnification for revealing subtle changes in the world," .

[9] William T. Freeman Frédo Durant Ce Liu, Antonio Torralba and Edward H. Adelson, "Motion magnification," .