ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-008

EDITE - ED 130

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Yingbo LI

le 27 Février 2012

CONSTRUCTION AUTOMATIQUE DE RESUMES MULTI-DOCUMENTS MULTIMEDIA

Directeur de thèse : **Bernard MERIALDO**

Jury

M. Georges LINARES, Professeur, LIA, Université d'Avignon	Président
Mme. Jenny BENOIS-PINEAU, Professeur, LaBRI, Université Bordeaux 1	Rapporteur
M. Nicu SEBE, Professeur, University of Trento	Rapporteur
M. Georges QUENOT, Directeur de recherches, LIG, Université Joseph Fourier	Examinateur
M. Gerhard RIGOLL, Professeur, MMK, Technische Universität München	Examinateur
M. Bernard MERIALDO, Professeur, EURECOM	Directeur de thèse

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

THÈSE

Thesis

AUTOMATIC CONSTRUCTION OF MULTI-DOCUMENT MULTIMEDIA SUMMARIES

Yingbo Li

Supervisor: Bernard Merialdo

February 27, 2012

This research was supported by EURECOM under the RPM2 (Multimedia, Multi-documents and Multi-opinions Summarization, http://www.rpm2.org) project from ANR (The French National Research Agency).

Dedicated to Yunxia for our sweet life

Acknowledgements

First, I wish to thank my advisor, Professor Bernard Merialdo, for his supervision during my Ph.D. study. Bernard is a great supervisor with abundant experience on research and teaching. Without his continuous support, I cannot finish this thesis. And I would like to thank the members of thesis committee for their helpful feedback and enthusiastic support.

Many friends in Eurecom, such as Simon Bozonnet, Moctar Mossi, Antitza Dantcheva, Xueliang Liu, Jinbang Chen, Xuran Zhao, Rui Min, Carmelo Velardo, Claudiu Tanase, Miriam Redi and others, also help me a lot during my research. They give me a lot of useful suggestion, and offer a lot of help to my experiments. Their help enabled me to finish my research successfully. As well, since I am a foreigner in France, these friends also help me a lot in my daily life. I will remember our happy life together in Sophia-Antipolis forever.

My graduation as a Ph.D. cannot be achieved without the support of my parents. In my life until now, I have met with a lot of troubles which seemed hard to be conquered, but at last I successfully resolved these troubles with their encouragement and support. I am very fortunate to live in such a great and warm family.

At last, I want to express my love and appreciation to Yunxia. She stayed in France more than two year together with me, and the life in a foreign country was actually not so easy. But Yunxia carefully took care of me in our daily life and I really felt life more meaningful with her encouragement. Every success in my life cannot be achieved without her effort. I really wish our small family is always healthy, sweet and blessed.

Abstract

With the increase of video quantity on the Internet, multimedia information processing has been a focused topic in the recent years. Among the techniques for multimedia information processing, video summarization has become an important tool. Some successful approaches have been proposed by the researchers in multimedia community. In this thesis, we propose our novel video summarization algorithm, Video-MMR (Video Maximal Marginal Relevance) based on visual information by mimicking MMR (Maximal Marginal Relevance) in text summarization. Video-MMR is a generic algorithm regardless of the video genre and suitable for summarizing both a single video and a set of videos. Besides Video-MMR as our basis, we also develop our approaches as following:

1. Visual information is always the most important compared to acoustic and textual information. So we overcome limits of Video-MMR and propose a refinement, Video-MMR2, by only exploiting visual information.

2. Since in a video, visual information is only one of several cues, more variants of Video-MMR using multimedia cues are proposed by exploiting multimedia information such as text or audio. We extend Video-MMR to AV-MMR (Audio Video Maximal Marginal Relevance), Balanced AV-MMR, OB-MMR (Optimized Balanced Audio Video Maximal Marginal Relevance) and TV-MMR (Text Video Maximal Marginal Relevance). These multimedia MMR algorithms are generic algorithms which outperform Video-MMR if we take into account the text and audio information in the video.

3. In addition to the summarization algorithms, we also optimize the presentation of video summaries, otherwise a good summary can be corrupted by a bad presentation. So we try optimizing a static summary containing keyframes and keywords by suggesting the number of frames and text grams, and dynamic summary composed of video segments by optimizing average duration of segments.

4. In the domain of video summarization, we need an evaluation measure for new approaches. Many current measures are based on human assessment, and the automatic evaluation method for video summaries is still an open problem. In this thesis we propose an approach, VERT (Video Evaluation by Relevant Threshold) mimicking the evaluation measures BLEU and ROUGE in text community to facilitate the automatic evaluation procedure with the help of only a few human assessments. We describe the details of all the approaches and present experimental results.

Therefore, a framework on video summarization is proposed, including an algorithm of video summarization using visual cue, its variants exploiting more multimedia cues, an optimization measure of summary presentation, and a new evaluation method of video summaries. It allows us to manage and browse multiple videos more efficiently.

(In French) Face à l'augmentation de la quantité de séquences vidéo sur Internet, le traitement de l'information multimédia est devenu un sujet de recherche grandissant ces dernières années. Parmi les différentes techniques utilisées on peut notamment citer le résumé vidéo qui est devenu un outil important. Plusieurs approches ont déjà été proposées avec succès par les chercheurs de la communauté multimédia. Dans cette thèse, nous proposons un nouvel algorithme de résumé vidéo, intitulé Vidéo-MMR (Video Maximal Marginal Relevance), et basé sur l'information visuelle imitant ainsi l'algorithme MMR (Maximal Marginal Relevance) utilisé dans le résumé automatique de texte. Le Vidéo-MMR est un algorithme générique applicable à tous types de vidéos et adapté à la fois au traitement d'une vidéo tout comme à un corpus entier. Outre le Vidéo-MMR, nous développerons aussi des extensions comme suit:

1. L'information visuelle restant toujours la plus importante comparée aux informations acoustiques et textuelles. Ainsi nous proposons de repousser les limites du Vidéo-MMR dans une version améliorée Vidéo-MMR2, grâce à une meilleure exploitation de l'information visuelle.

2. Etant donné que dans une vidéo l'information visuelle est seulement une composante parmi d'autres, plusieurs variantes de Vidéo-MMR utilisant différentes composantes multimédia sont proposées en exploitant des informations textuelles et audio. Nous étendons ainsi algorithme le Vidéo-MMR à AV-MMR (Audio Video Maximal Marginal Relevance), Balanced AV-MMR, OB-MMR (Optimized Balanced Audio Video Maximal Marginal Relevance) et TV-MMR (Text Video Maximal Marginal Relevance). Ces algorithmes MMR multi-média tiennent compte du texte et de l'information audio dans la séquence, et sont des algorithmes génériques qui donnent de meilleures performances que la version simple Vidéo-MMR.

3. En plus, des algorithmes de résumé, nous avons également optimisé la présentation des résumés vidéo, afin de prendre en compte les contraintes de présentation dans la construction du meilleur résumé. Nous avons optimisé un résumé statique contenant des images et des mots clés en suggérant le nombre d'images-clé et de segments de texte, ainsi qu'un résumé dynamique composé de segments vidéo en optimisant le choix de la durée moyenne des segments.

4. Dans le domaine des résumés vidéo, nous avons besoin d'une mesure d'évaluation des nouvelles approches. De nombreuses mesures actuelles sont basées sur l'évaluation manuelle, et la méthode d'évaluation automatique de résumé vidéo est toujours un problème d'actualité. Dans cette thèse, nous proposons une approche appelée VERT (Video Evaluation by Relevant Threshold) imitant les mesures d'évaluation BLEU ou ROUGE développées par la communauté de traitement des textes pour faciliter la procédure d'évaluation automatique avec l'aide de seulement quelques évaluations humaines. Nous décrivons en détails toutes ces approches et présentons des résultats expérimentaux.

Ainsi nous nous concentrons dans cette thèse sur le résumé vidéo, en particulier un algorithme de résumé vidéo utilisant des repères visuels, des variantes exploitant davantage de composantes multimédia, une optimisation de la présentation des résumés, et une nouvelle méthode d'évaluation de la qualité des résumé vidéo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid increase of the amount of videos on the Internet is obvious now. Every day many people upload and share news videos, personal videos and so on. How to manage such a large amount of visual data is a serious problem for human beings, so it is an active research topic nowadays. Video summarization has been identified as an important component to deal with video data. Research into video summarization produces an abbreviated form by extracting the most important and pertinent content in the video. Video summaries can be used in various applications, such as searching systems and interactive browsing, then the user can use it to manage and access digital video content [Lienhart et al., 1997] [Barbieri et al., 2003] [Money and Agius, 2007] [Lew et al., 2006].

Since the video is multimodal with the information of sound, music, still images, moving images and text [Dimitrova, 2004], video summarization is more complex than text summarization and other text processing techniques. In addition to the low-level features, video summarization needs to consider semantics in the video to facilitate the understanding of the summary.

While a lot of effort has been devoted to the summarization of a single video [Money and Agius, 2007] [Yahiaoui et al., 2001], less attention has been given to the summarization of a set of videos [Yahiaoui et al., 2001]. With the increase in quantity, it is more and more often that videos are organized into groups, for example, the YouTube website presents the related videos in the same webpage. Therefore the issue of creating a summary for a set of videos is getting an increased importance, which follows the trend that is now well established in the text document community. Multi-document video summarization has been extensively studied [Wactlar, 2001] [Dumont and Merialdo, 2008] [Das and Martins, 2007] [Wang and Merialdo, 2009]. However there are still many limitations. Many existing algorithms only consider the features from the video track, and neglect the audio track because of the difficulty of combining the information of audio and video. Several current algorithms [Sugano et al., 2002] [Furini and Ghini, 2006] [Xu et al., 2005] consider both audio track and video track, but they are domain specific. Motion features based on MPEG-7 and highlight detection by analyzing audio class and audio level are the main measures in [Sugano et al., 2002], while it is obvious that this algorithm is merely suitable for the video with the information of MPEG-7. In [Furini and Ghini, 2006] the authors consider that the video segments corresponding to silence in the audio track are useless, while there are cases where these segments may contain important information. In [Xu et al., 2005], the authors have invented an algorithm to summarize music videos: the authors detect the chorus in audio and the repeated shots in video track. Three successful algorithms above are examples of video summarization using both visual and audio information, but each of them only focuses on one domain. There are not many generic algorithms for video summarization until now, one reason being that in a domain-specific algorithm it is easier to utilize some special features or characteristics. For example, in sports video the shout of audience is a strong indication that the current visual information is likely to be important. In the generic algorithm, we cannot rely on these specific characteristics.

In this thesis, we propose our generic summarization algorithm for multiple videos, Video-MMR, which borrows the idea from MMR (Maximal Marginal Relevance) [Carbonell and Goldstein, 1998] for text summarization. Then we add audio information into Video-MMR and develop it to AV-MMR (Audio Video MMR), Balanced AV-MMR and OB-MMR (Optimized Balanced AV-MMR) step by step. By adding text information transcribed from speech, we propose TV-MMR (Text Video MMR). In addition, we suggest the optimal number of frames in the predefined stationary space containing frames and text by Video-MMR, and the optimal duration of segments for video skim by TV-MMR. Besides the above improvements on MMR by adding multimedia cues, we improve MMR by analyzing its characteristic and propose Video-MMR2 which only exploits visual information too but outperforms Video-MMR.

After generating video summaries, we need to evaluate their performances. Consequently, the evaluation of video summaries [Das and Martins, 2007] is a popular problem, still open to innovation. People can easily distinguish between ”good” and ”bad” summaries, but an ideal ”best” summary does not exist, so that it is difficult to define a quality measure that can be automatically computed. It is still possible to set up experiments involving human beings to evaluate video summaries, but these experiments are costly, time-consuming, and cannot easily be repeated, which impairs the development of many algorithms based on machine learning techniques. A good quality measure that can be automatically computed, and shows a strong correlation with human evaluation is therefore of great interest. Similar situations have already been encountered, for example in the domain of machine translation [Nilsson et al., 2007] or text summarization [Lin and Hovy, 2003], and the techniques proposed in these domains are used as a source for inspiration. Then our successful evaluation approach, VERT (Video Evaluation by Relevant Threshold) is proposed.

In general the research in this thesis makes the following contributions:

1. Multi-video summarization algorithms by visual information: Video-MMR and Video-MMR2.

2. Multi-video summarization algorithms by visual and acoustic information: AV-MMR, Balanced AV-MMR and OB-MMR.

3. Multi-video summarization algorithm by visual and textual information: TV-MMR.

4. The optimization of summary presentation for static summary and dynamic summary.

5. Automatic evaluation measure for video summaries: VERT.

This thesis is organized as follows: Chapter 2 reviews the state-of-the-art in the domain of video summarization. In Chapter 3 we review the visualization method of video summaries to illustrate the properties of the video, which inspires the research of our approach of video summarization, Video-MMR in the same chapter. Then we improve our Video-MMR by adding multimedia cues and propose AV-MMR, Balanced AV-MMR, OB-MMR and TV-MMR in Chapter 5. In Chapter 5 we also propose the approaches to optimize summary presentations for static summary and dynamic summary. Later we analyze characteristics of Video-MMR and evolve Video-MMR to Video-MMR2 in Chapter 4. Chapter 6 describes our evaluation approach for video summaries, VERT. At last we conclude this thesis with some final remarks in Chapter 7.

Chapter 2

The State of The Art

Video summarization has been an important tool to deal with digital video in the recent years. The task of video summarization is to generate a summary for one or more videos. The research on video summarization has been focused by the researchers. In this chapter, we will review the basic knowledge of video summary, the previous techniques of video summarization, the visualization of video summaries, the current evaluation measures and the application of video summarization.

2.1 The forms of video summary

Until now the forms of video summary can be clustered into two kinds [Truong and Venkatesh, 2007]: stationary images, also called keyframes and storyboard, and moving images, also called video skims. Fig. 2.1 illustrates the relations of keyframes, video skim and the original video.

As shown in Fig. 2.1, keyframes are a cluster of important frames which can represent and cover the most important content of the video. A good set of keyframes should enable the user to maximally understand the original visual content of the video.

In Fig. 2.1 we can also see that video skim is the concatenation of video segments, which may include their corresponding audio, extracted from the original video. A video skim uses much shorter video clips to represent the major information in the original video.

Video skim is able to include audio and motion in the summary, while keyframes can only represent stationary images (possibly with text sometimes in storyboard). However, keyframes can use shorter time and less space to review and present the same content compared with video skim.

Two forms of video summary are able to transform to each other. We can transform keyframes into video skim by adding fixed-size segments and audio [Truong and Venkatesh, 2007]. While, video skim can degenerate to keyframe by subsampling.

2.2 The techniques of video summarization

We will explore the useful features, underlying computation criterion and the current techniques of video summarization in the section.

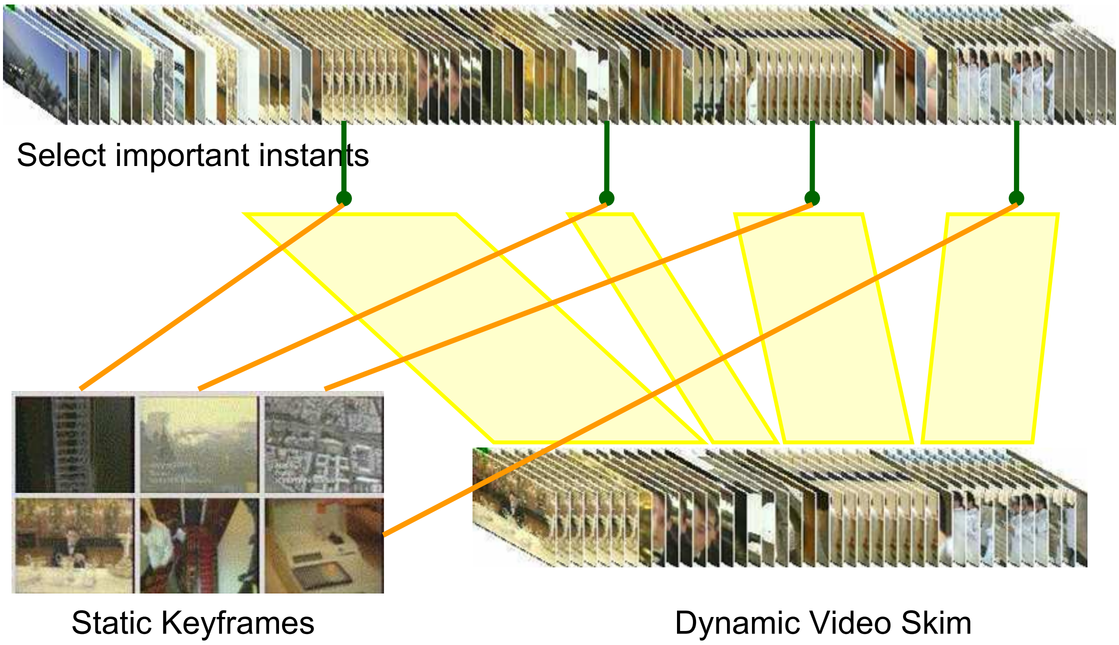


Figure 2.1: Two forms of video summary

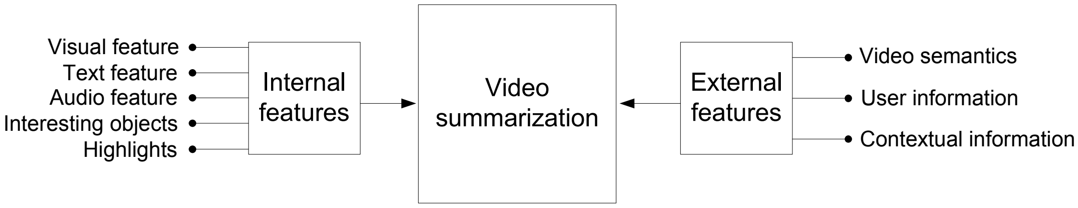


Figure 2.2: The features useful in video summarization

2.2.1 The multimedia features

The useful features of the video for video summarization can be divided into internal features and external features, shown in Fig. 2.2. The internal features of the video are the information which can be extracted from the video itself. The internal features include:

1. Visual feature. The color features like HSV histogram of the keyframe, texture features, SIFT features and the spatial arrangement are usually exploited.

2. Text feature. The text from the video can be got from video frames and transcribed from audio channel of the video. The text in video frames can be captured by Optical Character Recognition (OCR) techniques [IMPEDOVO et al., 1991]. And the text from the speech of the video is usually transcribed by the techniques of Acoustic Speech Recognition (ASR) [Cole, 1997], like in [Li et al., 2011].

3. Audio feature. The most common audio feature is Mel Frequency Cepstral Coefficients (MFCCs), as [Li and Merialdo, 2010b] and [Gao et al., 2009].

4. Interesting objects. Since video summarization is a human-oriented task, the face is a kind of interesting objects in the processing [Li and Merialdo, 2012]. While in the soccer video, the soccer is a definitely important feature [Ekin et al., 2003].

5. Highlights. The highlights may vary according to the genre of the video. But the fast motion between video frames and big change of the audio are usually considered as the interesting highlights.

The external features of the video, which are information outside the video itself, are:

1. Video semantics. The event metadata in MPEG-7 format [Jaimes et al., 2002] and sound-source location [Erol et al., 2003] are two examples of this kind of feature.

2. User information. The user information can be obtained by recording the responses, interaction and preference of the user to the specific videos [Money and Agius, 2007]. Normally the user information is categorized into the obtrusively sourced information from the conscious user providing more necessary detail and unobtrusively sourced information from the unconscious user. The annotation to the video in social media is a typical obtrusively sourced information.

3. Contextual information. In [Silva et al., 2005] the authors domain-specifically summarize videos by the contextual information of human movement during the recording by the camera.

2.2.2 Underlying computation criterion

No matter which form of video summary we want to get, keyframes or video skim, the underlying computation criterion by using the features mentioned in Section 2.2.1 would not change. An ideal video summary should eliminate the redundancy in the original video, minimize the correlation between keyframes or video segments in the summary, cover major information of the video, and keep the important events from the video, shown in Fig. 2.3.

- Eliminate the redundancy. A video contains a lot of unnecessary frames or segments which include the duplicated information, therefore it is meaningful to eliminate the redundancy in the original video and create video summary much shorter than the original video. We can use methods of selecting frames or segments by maximizing the average similarity to the original video [Cooper and foote, 2002], clustering [Hanjalic and Zhang, 1999], dynamic programming [Lu et al., 2004a] [Lu et al., 2004b] and so on.

- Minimize the correlation. If keyframes or video segments in video summary are of high correlation, it means that keyframes or video segments are duplicated and need more keyframes or video segments to represent the same video content. Therefore during the summarization, it is necessary to prove the most minimum correlation between keyframes or video segments. This procedure can be formulated as

$$\{r_1, r_2, \ldots, r_k\} = \arg\min_{r_i} Corr(f_{r_1}, f_{r_2}, \ldots, f_{r_k}) \tag{2.1}$$

  where $Corr(\cdot)$ is a correlation measure. Furthermore this computation can consider frames/segments pairs or successive frames/segments [Truong and Venkatesh, 2007]. In [Doulamis et al., 1998] it uses the correlation value of frame pairs of videos, like Eq. 2.2.

$$Corr(f_{r_1}, f_{r_2}, \ldots, f_{r_k}) = (\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Corr(f_{r_i}, f_{r_j})^2)^{1/2} \tag{2.2}$$

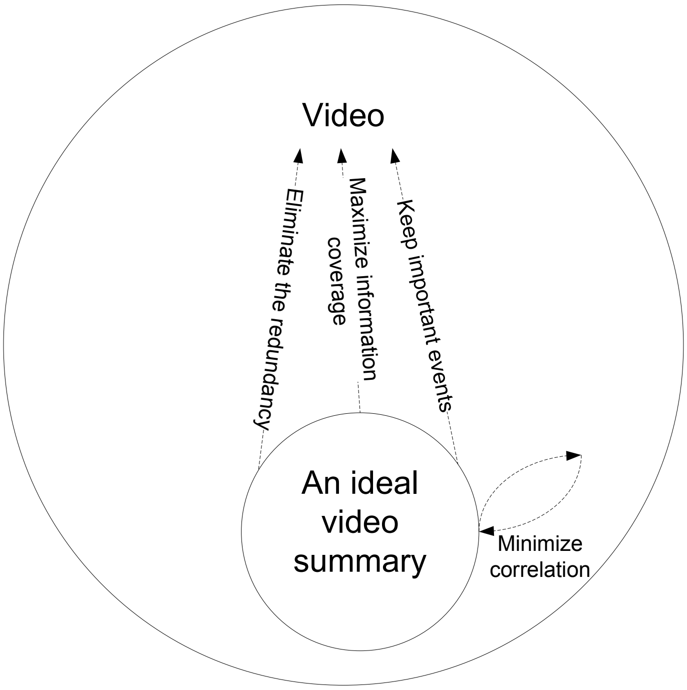


Figure 2.3: Underlying computation criterions

where $Corr(f_{r_i}, f_{r_j})$ is the correlation of f_i and f_j. When only the correlations of successive elements are considers, Eq. 2.1 is reformulated as

$$\{r_1, r_2, \ldots, r_k\} = \arg\min_{r_i}\{\sum_{i=1}^{k-1} Corr(f_{r_i}, f_{r_{i+i}})\} \tag{2.3}$$

- Maximize the information coverage. The aim of video summary is to maximally represent the video content of the original video. In [Chang et al., 1999] it employs a greedy method which extracts the frame with maximum coverage at each step. And [Li and Merialdo, 2010e] proposes an incremental procedure to produce video summary by maximizing the coverage of video content and minimizing frames' correlation.

- Keep the important events. Though a video summary covering the major video content is ideal, it is also necessary to keep the important events in the summary. Especially in video summarization of sports video it is a popular criterion, so it is reasonable to construct the library of events, such as "catch", "shoot" and so on and keep these important frames or segments in the summary. Furthermore, highlights can be detected by combining the information of audio, visual, text and cinematic information. In [Rui et al., 2000] it is assumed that the announcer's excited speech is related to exciting baseball segments. In addition, [Radhakrishnan et al., 2004] indicates that the detection of outlier events with a statistical model is a way for event detection.

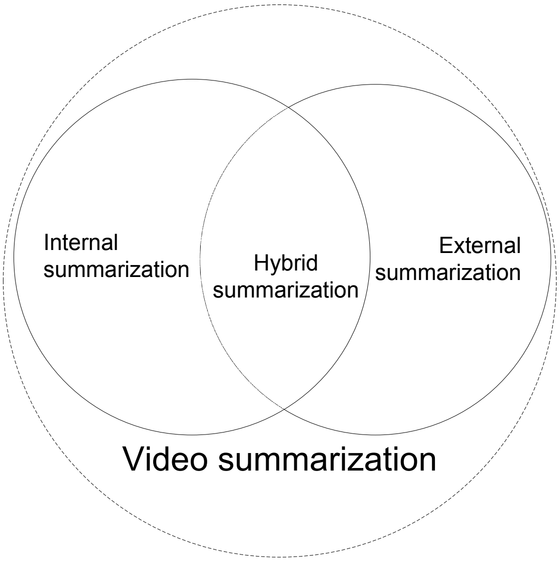


Figure 2.4: The categories of the techniques of video summarization

2.2.3 Current techniques of video summarization

According to the kinds of the features exploited in video summarization, we can divide the current techniques of video summarization into internal summarization, external summarization and hybrid summarization [Money and Agius, 2007], as shown in Fig. 2.4.

2.2.3.1 Internal summarization

Internal summarization, the most popular kind of approaches, uses the information from the image, audio and text in the video.

Image features includes the color, texture, shape and the objects like face. We can use image features to segment the video into shots by big change between frames and compute the similarity between frames. Image features are exploited in [Wang and Merialdo, 2009], in which the authors propose a non-domain specific algorithm by identifying representative keyframes. While there are many domain-specific approaches by image features too, such as the techniques for sports video in [Ekin et al., 2003] and [Shih and Huang, 2005]. In [Calic et al., 2007] the authors exploit universal and intuitive rules of comic-like videos by dynamic programming. Since specific objects or events are helpful to the domain-specific video summarization, they are popularly exploited. TRECVID BBC Rushes summarization [Smeaton et al., 2006] is only for the videos of BBC Rushes, for example a successful framework proposed by [Naci et al., 2008]. Another successful summarization algorithm for Rushes is based on the unsupervised classification [Rossi et al., 2009], which is evaluated by comparing the results for TRECVID data and a manually annotated ground truth.

Audio features in the video include MFCC, energy and audio genres including speech, music, noise, silence and other special sounds. The audio is useful in segmenting the video. Moreover, the audio is exploited in many domain-specific methods, especially for sports video [Xu et al., 2003] [Otsuka et al., 2005] [Cai et al., 2003].

Many techniques exploit both visual and audio information to summarize the video. An example is the algorithm in [Furini and Ghini, 2006] which summarizes video content based on the segments of noise or silence in the audio. Another successful example [Ma et al., 2005] is to use user attention to image and audio features to probabilistically summarize video. Many of these approaches are domain specific.

Text in the video includes the separated subtitle and the ”burnt in” caption. Text features usually contain detailed information of the video content, for example actor names, movie titles, and the location. The caption in the video can be captured by OCR (optical character recognition). [Zhang and Chang, 2002] exploits image, audio and text information in its domain-specific summarization algorithm for baseball video. And most approaches until now exploiting text are domain-specific. Besides above two sources, the text transcribed from audio channel can represent both textual and acoustic information, and improve the summarization, like the algorithm TV-MMR [Li et al., 2011].

In this thesis all proposed algorithms of video summarization belong to internal summarization.

2.2.3.2 External summarization

External summarization techniques analyze information outside the video itself. As mentioned in Section 2.2.1, video summarization can use the information of video semantics, user information and contextual information. External summarization is rare now, and external information is mostly exploited in hybrid summarization.

2.2.3.3 Hybrid summarization

Hybrid summarization techniques exploit both internal and external information and can be the combination of internal and external summarization algorithms. Hybrid summarization maximizes the advantage of each approach in the combination and minimizes the disadvantage. Some domain-specific algorithms for sports video have exploited hybrid techniques [Coldefy et al., 2004] [Fayzullin et al., 2004]. For non-domain specific approaches, hybrid method is useful too [Yu et al., 2003], where hybrid technique reduces manual work through segmenting videos by internal analysis. The user information used in hybrid summarization including user behavior such as glance, gaze, walk, and jump can be recorded during the record [Coldefy et al., 2004] [Fayzullin et al., 2004], watching the video [Shipman et al., 2003]and both [Lin and Tseng, 2005] by wearable, sunglasses and headband cameras. Furthermore, Internet is popularly used to collect user information [Lienhart, 2000].

2.3 The size of video summary

The decision of the size of video summary, which is the number of frames for keyframes or time duration for video skim, is still an open problem. The more the number of keyframes is or the longer time duration is, the better the coverage of video content is. However, more frame number or longer video skim corrupts a good summary, increases the redundancy in the summary and wastes the user longer time to watch the summary.

The current techniques normally decide the size of video summary before or during the procedure of video summarization:

1. Before video summarization, summary size is usually manually decided.

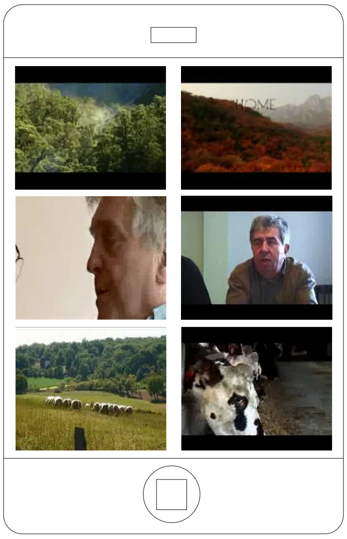


Figure 2.5: An example of video summary displayed on mobile phone

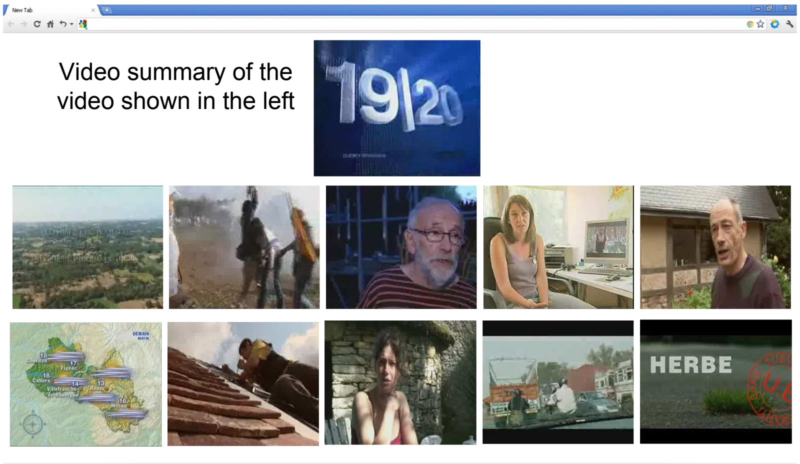


Figure 2.6: An example of video summary displayed on computer monitor

- We first talk about the summary size for keyframes. For the display on mobile device, a common size of video summary should be 6 like Fig. 2.5, while for the human assessment on computer monitor for video summary, the number between 10 and 20 is more appropriate as shown in Fig. 2.6. More keyframes in a limited space and time are out of the understanding ability of human beings. However a summary with a predefined size cannot promise to present all the important contents of the video. So in [Hammoud and Mohr, 2000] the authors propose to decide the summary size during the clustering procedure.
- For video skim, the duration of video skim highly depends on the dynamics in the video. But normally it is hard for a person to carefully remember the video content longer than one minute.

2. To determine the summary size during the summarization for both keyframes and video skim, we need a criterion such as visual change, similarity to the original video, or something else of the produced video summary, to stop the procedure of video

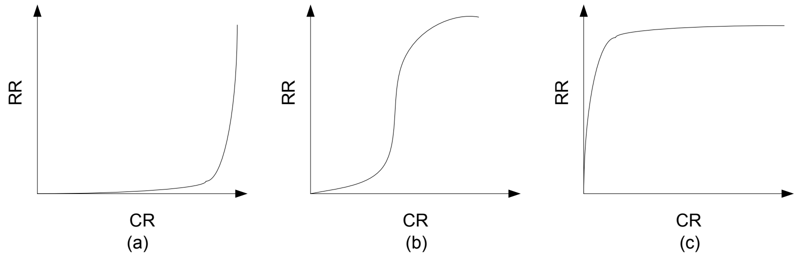


Figure 2.7: Different types of video summarization

summarization. Therefore, a more dynamic video needs a summary with a larger size to represent video contents.

2.4 The evaluation of video summary

2.4.1 General summary quality criteria

Before proposing a good evaluation method, it is necessary to define the property of a good video summary. In general, a good video summary follows these characteristics:

- The summary should be as short as possible compared with the original video,
- The important information must be represented by the summary,

However, it is impossible to satisfy both situations, because they are contrary. Therefore, we need a compromise. Assume that the summary S is from the video F, then the first two characteristics can be formulated as [Taskiran, 2006]:

$$Compression\ Ratio(CR) = \frac{length\ of\ S}{length\ of\ F} \tag{2.4}$$

$$Retention\ Ratio(RR) = \frac{information\ in\ S}{information\ in\ F} \tag{2.5}$$

A good summary should have low CR and high RR, which is the form illustrated in Fig. 2.7(c) [Taskiran, 2006].

2.4.2 Current evaluation measures

There is not an ideal summary for a video, so it is difficult to get perfect objective and quantitative evaluation measures. Every person has his own view for the video content. But for a given summary, most person may make coherent evaluation results. However, human evaluation is hard to implement, time-consuming and not automatic.

One popular evaluation method until now is the quiz method, which is to request the user to answer prepared questions about clearness, information coverage and so on. The quiz score increases when the user is more satisfied with the summary. In [He et al., 1999] the authors give many users the quiz about the reflection of the key ideas in video skim after they watch the original video and its video skim. In [Taskiran et al., 2002] the quiz is a little different, and the method is to check if the user can find original video segment where the face in video skim is included, and match video skim with text and frame from original video. Currently we can use a crowdsourcing platform such as Amazon's

Mechanical Turk [1] [Kittur et al., 2008] to collect the selections and quiz answers from various users [Rudinac et al., 2011] [Hanjalic and Larson, 2011].

Another popular kind of summary evaluation now is based on the number of included events in the summary. In [Ferman and Tekalp, 2003] the quality of video summary is related to the number of redundant or missing keyframes. As well, precision and recall values are a good way too. In [Ekin et al., 2003] the authors decide the summary evaluation by the precision and recall values of goal, referee, and penalty box detection.

2.4.3 The evaluation benchmark of TRECVID

NIST TRECVID (TREC Video Retrieval Evaluation) [Smeaton et al., 2006] has offered an evaluation benchmark in the community of multimedia retrieval in the past several years. TRECVID has successfully been facilitating the content-based analysis and retrieval of the video. TRECVID involves a lot of professional researchers to construct the ground truth by manually annotating, tagging and selecting their best summaries for videos. So the ground truth in TRECVID is the precious data to evaluate different algorithms.

In 2007 and 2008 the TRECVID summarization task dealt with BBC rushes videos, and created the summary which uses a minimal number of frames to optimally present the information of the summary and facilitating objects/event recognition. The system task in rushes summarization was to automatically create an MPEG-1 summary clip less than or equal to a given duration showing the main objects and events in rushes videos.

The evaluation of automatically generated summaries from the participants in TRECVID summarization task depended on both objective and subjective measures [Over et al., 2007] [Over et al., 2008].

- Subjective methods depended on the importance segments, redundancy and the temp/rhythm in video summary. Before the evaluation, 5 retired adults with computer skills created many important items such as a person, thing or event which could identify a video segment different from the others. During the evaluation 10 assessors evaluated the video summary depending on the found items in the automatic summaries. If more items such as objects and events were found, the corresponding automatic summary was better. Besides that, the assessors also should give their opinions on the following statements:

  1. ”This summary contains many color bars, clapboards, all black or all white frames.”
  2. ”This summary contains many nearly identical segments.”
  3. ”This summary is presented in a pleasant tempo and rhythm.”

- While, objective measures were the following: (1) percentage of desired segments found as judged by assessor; (2) presence of junk (color bars, clapboards, empty frames), as judged by the assessor; (3) amount of near redundancy, as judged by assessor; (4) satisfaction with tempo and rhythm of presentation, as judged by the assessor; (5) assessor time taken to determine presence/absence of desired segments; (6) duration of summary relative to the 2% duration target; (7) elapsed time for summary creation.

[1] http://www.mturk.com

The success of this open and systematic evaluation is based on the following factors [Truong and Venkatesh, 2007]:

1. A common video set. Only researchers use the same database of videos, such as TRECVID and Open-video Archive, and then they can compare their own summaries for the same video. In addition, it is easy to collect the contribution from diverse people.

2. Ground Truth. Different people have different opinions of the same summary, but it is promising to get the ground truth after collecting results from a large amount of people, because results would converge. We can consider these coherent results as ground truth for a video summary and use them to measure the quality of the other summaries for the same video. Furthermore, since TRECVID is a benchmark, different algorithms of video summarization is evaluated in the same evaluation perspective, so different algorithms can be compared.

2.5 The visualization of video summary

Since there are two forms of video summary, keyframes and video skim, the visualization measures of video summary are for these two types.

The visualization of keyframes can be storyboard, like Fig. 2.8 [Christel, 2006], and dynamic slideshow, like 2.9 [Chen et al., 2010]. Some research works [Komlodi and Marchionini, 1998] have shown that users prefer storyboard to dynamic slideshow. The drawbacks of storyboard are that it is hard to capture all the details and the spatial relation of keyframes. Therefore there are some alternative methods of storyboard:

- Some alternative methods of keyframe visualization have been proposed by adding the semantic. One example is proposed by [Li et al., 2009]. In [Li et al., 2009] the authors present all the keyframes in a circle and the positions of keyframes are decided by their similarities, and an example is shown in Fig. 3.3.

- Some other visualization methods present the frame with different sizes corresponding to their importance. The frames with larger sizes mean more important than the frames with smaller sizes. ”Video Manga” [Uchihashi et al., 1999] shown in Fig. 2.10 is a successful example in this kind.

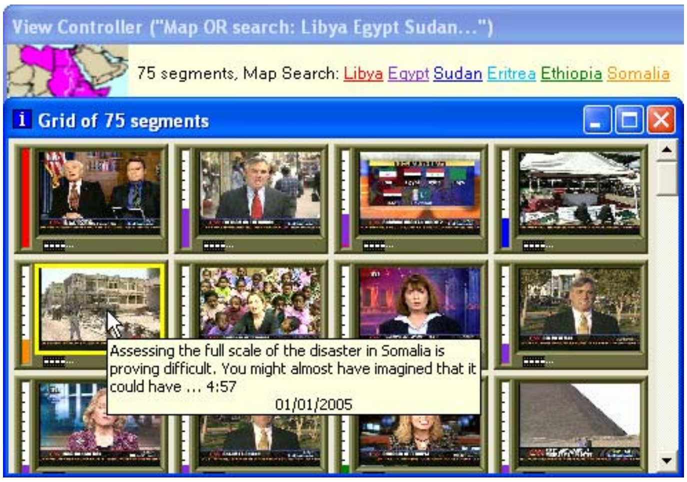
While for video skim, the most popular visualization measure is the concatenated short video by video segments of video skim according to time order.

2.6 The application of video summarization

The application of video summary can be categorized into the following kinds [Money and Agius, 2007] [Truong and Venkatesh, 2007]:

1. Browsing. The video summary can help the user quickly find the interesting video in a large amount of videos.

2. Content navigation. Traditionally the user needs to playback the video again and again to check the interesting shot. But it is time-consuming and hard to control. While video summaries enable the user to easily access the important content of the



View Controller ("Map OR search: Libya Egypt Sudan...")

75 segments, Map Search: Libya Egypt Sudan Eritrea Ethiopia Somalia

Grid of 75 segments

Assessing the full scale of the disaster in Somalia is proving difficult. You might almost have imagined that it could have ... 4:57

01/01/2005

Figure 2.8: An example of storyboard

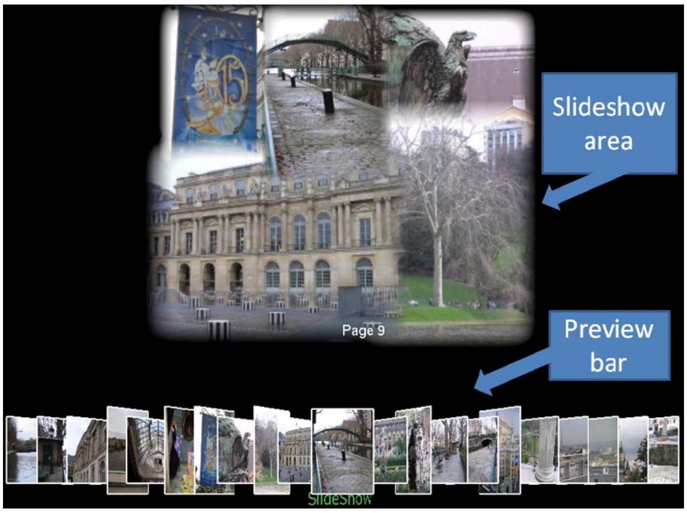


Figure 2.9: An example of slideshow

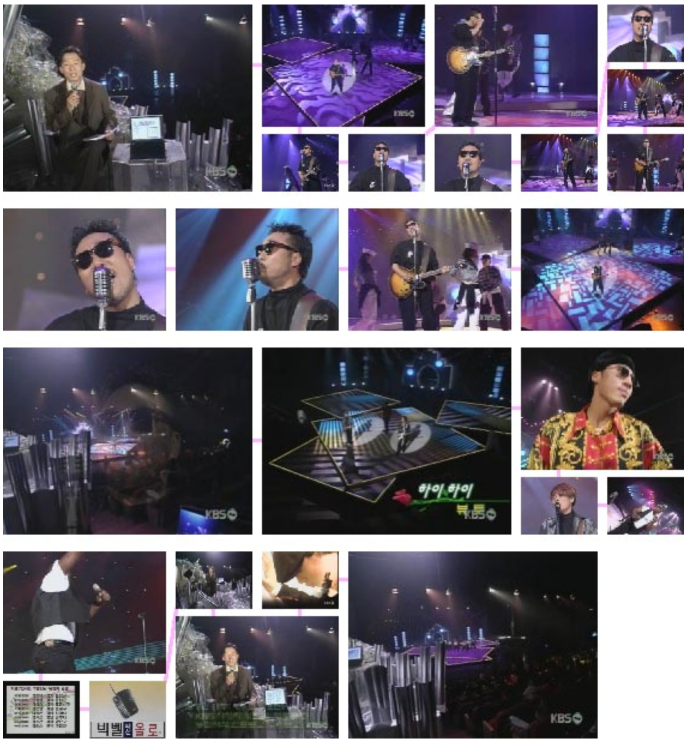
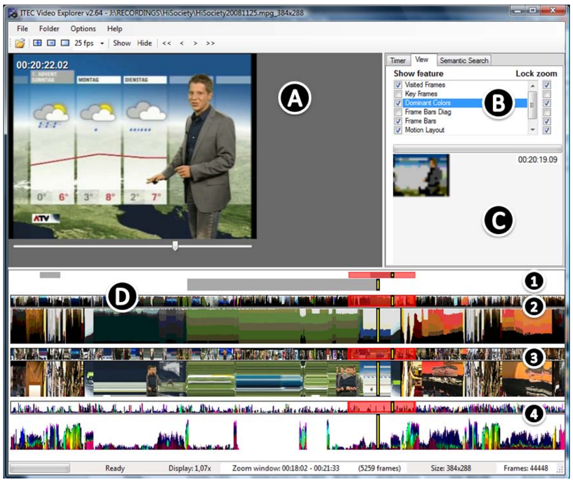


Figure 2.10: An example of Video Manga


ITEC Video Explorer v2.64 - J:\RECORDINGS\HiSociety\HiSociety20081125.mpg_384x288

00:20:22.02

Show feature | Lock zoom
☑ Visited Frames ☑
□ Key Frames □
☑ Dominant Colors ☑
□ Frame Bars Diag □
☑ Frame Bars ☑
☑ Motion Layout ☑

00:20:19.09

Ready | Display: 1,07x | Zoom window: 00:18:02 - 00:21:33 | (5259 frames) | Size: 384x288 | Frames: 44448

Figure 2.11: An example of video navigation

whole video in a short time. An example is shown in Fig. 2.11 [Schoeffmann and Boeszoermenyi, 2009].

3. Content analysis. Video summarization extracts the most interesting highlights and events from the video, which shortens the computation time of the future processing like video search, because video search can directly deal with video summary which is of less but concentrated information.

Chapter 3

Video-MMR

In this chapter we first discuss the relation of video frames from multi-video and reveal the possibility to construct video summary by the similarities between frames. Then we borrow the idea of text summarization based on text similarities and introduce it into the domain of video summarization.

3.1 Visualization of Multi-video summaries

In this section, we will present three visualization approaches of video summaries, including panel representation, circle representation and hierarchical representation [Li et al., 2009]. The redundancy in a video or a video set is shown in three visualizations, which illustrates the necessity of video summarization. Furthermore, from the visualized figures we can intuitively find the correlations of the visual and acoustic frames based on visual and acoustic features, so that we can use visual and acoustic features of video frames to quantify the relations between frames, between videos and between a video and a summary. Consequently, the visualization shows that it is able to summarize the video by computing the similarities and correlations between the features of video frames.

3.1.1 Panel representations

Most of the existing work on video summarization has dealt with the problem of a single video. In this thesis we are dealing with the issue of multi-video summarization, because it arises naturally in our daily application. Our experimental videos are from the partner of our project, the Wikio web site (http://www.wikio.fr/) - an Internet news aggregator which gathers news items, (both texts and videos) from a large variety of sources, and organizes them by articles on specific topics. People can therefore have a global view of an event as presented through different channels and with various comments. Often, a group of items comes with attached videos. Those videos can be of diverse type, recordings of TV programs, homemade videos, advertising clips, etc.

Our first activity is to design a mechanism to visualize the content of a set of videos, have a global overview of their content and detect possible similarities. Simple mechanisms like displaying a panel with some keyframes as Fig. 3.1 are very limited, because they are missing a lot of information.

A panel of video traces (a column of pixels extracted from each keyframe) provides a visualization that is more precise in the temporal domain, but less precise in the spatial domain. Fig. 3.2 shows such an example.

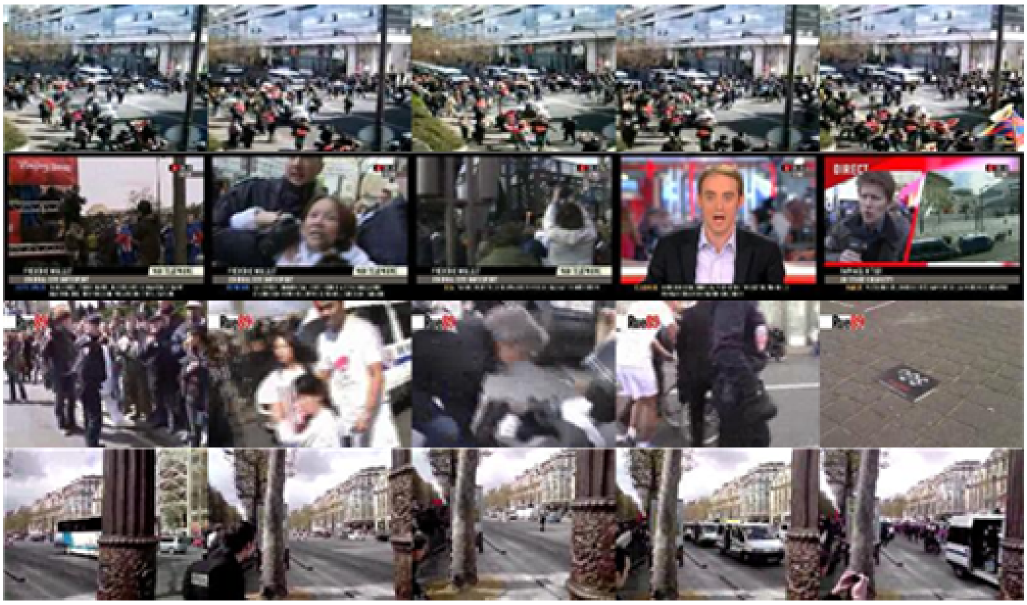


Figure 3.1: Keyframes from several videos

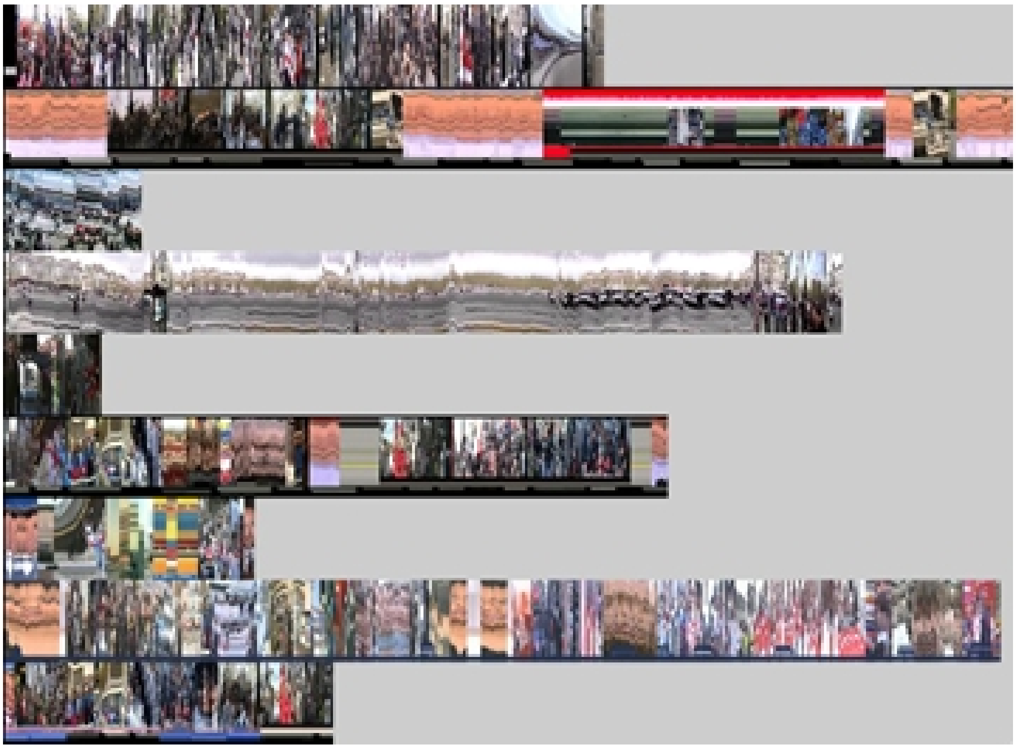


Figure 3.2: Panel of video trace images

3.1.2 Circle representations

To better represent the relationships between different videos, we have defined a circle space where multi-video frames can be positioned according to their visual content.

Assume that we have a group of videos, $\{V_1 \dots V_i \dots V_n\}$. For all the videos, we can compute the similarity value between every video and every frame of this group. It is described as following:

$$SIM(f, V_i) = \max_{k \in V_i} \{SIM(f, k)\} \tag{3.1}$$

where k is a frame from video group $\{V_1 \dots V_i \dots V_n\}$, and V_i is a video in this group. So the similarity value between frame f and video V_i, $SIM(f, V_i)$, is considered as the maximum similarity value between frame f and every frame k of video V_i. For example, the similarity value can be computed as the cosine between the feature vectors of frame f and k:

$$SIM(f, k) = \cos(X, Y) = \frac{X \cdot Y}{\| X \| \| Y \|} \tag{3.2}$$

where X is the feature vector of frame f, and Y is the feature vector of video frame k. A circle is used to visualize the relations of videos and frames. The videos are regularly placed on the edge of the circle. We just select one frame per second, kept in all the experiments of this thesis, which are shown as points inside the circle. The more similar a frame is to a video, the closer the point of this frame is to the point of the corresponding video. The coordinates of frame points in the circle are computed as:

$$P_f = \sum_{i=1}^{n} P_m \cdot SIM(f, V_i) \tag{3.3}$$

where P_f is the position of the frame f, P_i is the position of video i.

An example is in Fig. 3.3. The same color represents the frames from the same video. We can see that the special frames from two upper right videos will be far from circle center, but the frames of the other four similar videos are close to circle center and each other. At the same time, most points of the same color are close. From above analysis, we could conclude that the frames from the same video are clustered together and similar with each other, except that two different videos own the similar view contents. This allows to have a global view of the diversity of the visual content of a multi-video set.

3.1.3 Hierarchical representations

Now one frame per second is selected from the video group, $\{V_1 \dots V_m \dots V_n\}$. Initially one frame is one cluster. After that, at each step of clustering, the number of clusters is reduced one by one by merging the two closest clusters. In this way we could construct a clustering structure through the iterations of merging clusters, as Fig. 3.4. The distance between two frames is the Euclidean distance of HSV histogram, like Eq. 3.4. The distances between clusters are the average distances of all the possible frame pairs in the clusters, like Eq. 3.5.

$$DIS(f, s) = \sqrt{\sum_{i=1}^{j} (X_i - Y_i)^2} \tag{3.4}$$

$$DIS_{cluster} = \frac{\sum (DIS_{each\ pair})}{number\ of\ pairs} \tag{3.5}$$

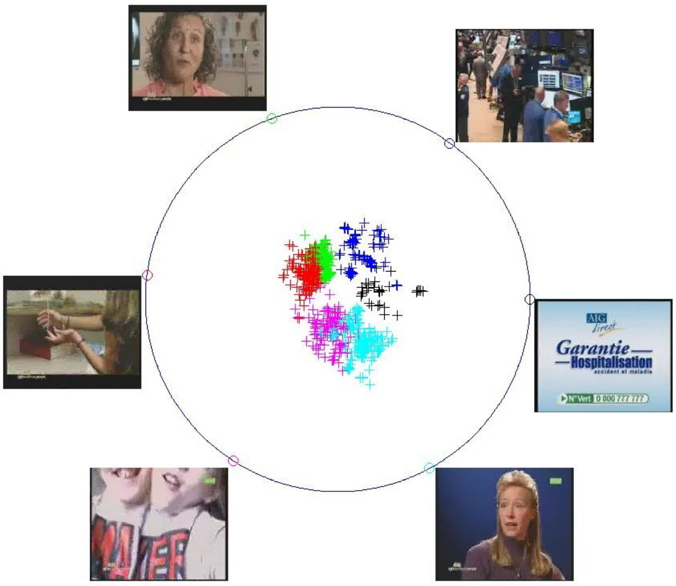


Figure 3.3: Example of circle representation for the visual content of a set of videos

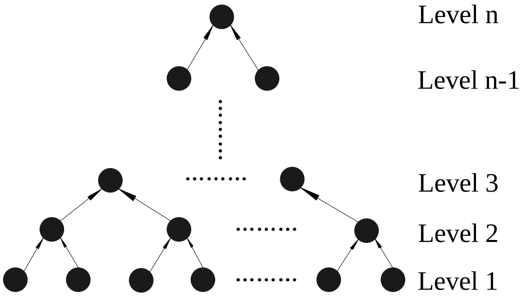


Figure 3.4: Hierarchical Level

where f and s are frames from video group. X is HSV histogram of frame f, and Y is HSV histogram of frame s. The argument j is the total number of tonal variations of histogram. Assume that we want to select a definite number of frames as video summary, and if the clusters' number of a level is just smaller than this number, the middle frames in the selected clusters will be the keyframes in final video summary. For example, in Fig. 3.4 the cluster number of level 3 is 30, which is just smaller than the required number 31, then the middle frames of 30 clusters in level 3 are the final video summary.

Fig. 3.5 is an example. Here we illustrate the procedure of selecting 5% frames as video summary one cluster by one cluster. This figure is explained as following: Fig. 3.5(a) in the first row is the middle columns of each frame per second of all the videos by time order. Fig. 3.5(b) is the clusters of frames and one color represents one cluster of frames. Fig. 3.5(c) shows that the selected clusters disappear, the middle frames of which appear in Fig. 3.5(d). Fig. 3.5(d) shows a simple summary. Hierarchical representation is useful in video content navigation.

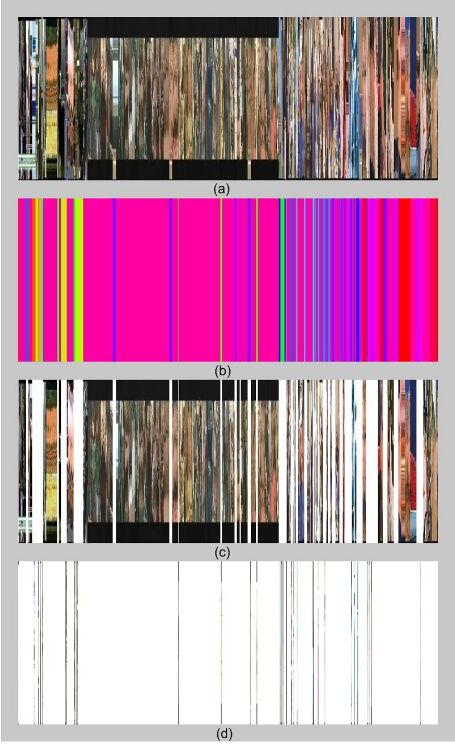


Figure 3.5: Example of hierarchical visualization

3.1.4 Conclusion of visualization

The above visualizations of the summaries from simple summarization approaches show that video frames are dependent and related by their similarities. The frames in a video and the frames from two videos with correlated contents are commonly more correlated. Therefore, this knowledge gives us the possibility and motivation to develop better video summarization algorithm.

3.2 Video Maximal Marginal Relevance (Video-MMR)

Various video summarization techniques have already been proposed. In this thesis, we propose a novel and effective approach for multi-video summarization: Video Maximal Marginal Relevance (Video-MMR), which extends a classical algorithm of text summarization, Maximal Marginal Relevance (MMR). Video-MMR rewards relevant keyframes and penalizes redundant keyframes, as MMR does with text fragments. Two variants of Video-MMR are suggested.

This section is organized as follows: Section 3.2.1 briefly reviews text summarization and presents the original MMR of text summarization. The theory of Video-MMR is in Section 3.2.2. Section 3.2.3 presents the experiments: we use a criterion of minimum distance with the original video to experimentally select the best variant of Video-MMR; we compare Global and Individual Summarization, and conclude that Global Summarization is better; meanwhile, summary quality of Video-MMR is assessed with human-generated ground truth, and compared with K-means algorithm. At last, we conclude this section with some final remarks.

3.2.1 Text summarization and MMR

The demand of quick information search is increasing. Traditional algorithms look for maximizing relevance to the user query. However, if the potentially relevant information is vast, tremendous redundant and duplicated documents are found. Consequently, it is necessary to consider the anti-redundancy property of document summarization. In the community of Natural Language Processing (NLP), automatic text summarization has been addressed by various algorithms over the past half century. A text summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that" in [Radev et al., 2002], where several summarization terms are explained too: "extraction" is a processing to identify and reproduce the important section verbatim; the task of "abstraction" is to generate the significant material by a different way; "fusion" coherently concatenates the extracted text; and "compression" is to delete the unimportant text from the original file. According to the above terms, we can define several types of the summarization [Das and Martins, 2007]: "extractive summarization" depends on sentences to produce the content of the summary; "abstractive summarization" produces a grammatical summary with the help of advanced language generation techniques; while in information retrieval (IR), "topic-driven summarization" depends on the preference of the user and focuses on a particular topic. For single-document summarization, the current algorithms of text summarization use the techniques of Naive-bayes, decision trees, Hidden Markov Models, log-linear model and neural networks and other special feathers according to the language. In the current approaches [Das and Martins, 2007] for multi-document summarization, there are several successful types: abstraction and information fusion, Maximal Marginal Relevance (MMR), graph spreading activation, centroid-based approach and multilingual approach.

Until now a popular and efficient one is MMR proposed by [Carbonell and Goldstein, 1998]. The Marginal Relevance (MR) of a document D_i with respect to a query Q and a selection of documents, S, is defined by the equation:

$$MR(D_i) = \lambda Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \tag{3.6}$$

where Q is a query or user profile, and D_i and D_j are text documents in a ranked list of documents R. D_i is a candidate in the list of unselected documents $R \backslash S$, while D_j is an already selected document in S. In the equation, the first term favors documents that are relevant to the topic, while the second will encourage documents which contain novel information not yet selected. The parameter λ controls the proportion between query relevance and information novelty. MR can be used to construct multi-document summaries by considering the set of all documents as the query Q, R as a set of text fragments, and iteratively selecting the text fragment D_{MMR} that maximizes the MR with the current summary:

$$D_{MMR} = \arg\max_{D_i \in R \backslash S} MR(D_i) \tag{3.7}$$

In [Carbonell and Goldstein, 1998], the authors indicate that MMR works better for longer documents and is extremely useful in extraction of passages from multiple documents for the same topics, when we consider the document passages as summary candidates. Since news stories contain a lot of repetition, the authors show that the top 10 passages contain a significant repetition by previous methods, while MMR reduces or even eliminates such redundancy.

3.2.2 Video-MMR

The goal of video summarization is to identify a small number of keyframes or video segments which contain as much information as possible from the original video. Video segments can be characterized by one or several keyframes, so we focus here on the selection of relevant keyframes. Because we want each frame can commonly represent a part of visual content in the original video, each summary frame ideally is of the same or very similar frames in the video except summary frames and the ideal distance between the summary and the video is 0. Consequently the quality of the summary S is measured by average distance between frames of the summary, S, and the other frames in the original video, V:

$$d(S,V) = \frac{1}{n}\sum_{j=1}^{n} \min_{f_j \in V\backslash S, g \in S}[1 - sim(f_j, g)] \tag{3.8}$$

where n is the number of frames in V. g and f_j are frames respectively from S and V. With this presentation, the best summary $\hat{S}$ (for a given length) is the one that achieves the minimum distance:

$$\hat{S} = \arg\min_{S}[d(S,V)] \tag{3.9}$$

Because video summarization has similarities with text summarization, we propose to adapt the MMR criteria to design a new algorithm, Video Maximal Marginal Relevance (Video-MMR) [Li and Merialdo, 2010e], for multi-video summarization.

When iteratively selecting keyframes to construct a summary, we would like to choose a keyframe whose visual content is similar to the content of the videos, but at the same time, which is different from the frames already selected in the summary, which is illustrated in Fig. 3.6. By analogy with the MMR algorithm, we define Video Marginal Relevance (Video-MR) by:

$$\textit{Video-MR}_S(f_i) = \lambda Sim_1(f_i, V\backslash S) - (1-\lambda)\max_{g \in S} Sim_2(f_i, g) \tag{3.10}$$

where V is the set of all frames in all videos, S is the current set of selected frames for the summary, g is a frame in S and f_i is a candidate frame for selection. Based on this measure, a summary S_{k+1} can be constructed by iteratively selecting the keyframe with Video Maximal Marginal Relevance (Video-MMR):

$$S_{k+1} = S_k \bigcup \arg\max_{f_i \in V\backslash S_k}\{\lambda Sim_1(f_i, V\backslash(S_k \bigcup f_i)) - (1-\lambda)\max_{g \in S_k} Sim_2(f_i, g)\} \tag{3.11}$$

Sim_2 is just the similarity $sim(f_i, g)$ between frames f_i and g. We need to define $Sim_1(f_i, V\backslash S_k)$. We consider two variants for this measure:

- The average similarity is the arithmetic sum:

$$AM(f_i, V\backslash(S_k \bigcup f_i)) = \frac{1}{|V\backslash(S_k \bigcup f_i)|}\sum_{f_j \in V\backslash(S_k \bigcup f_i)} sim(f_i, f_j) \tag{3.12}$$

- The average similarity is the geometric sum:

$$GM(f_i, V\backslash(S_k \bigcup f_i)) = [\prod_{f_j \in V\backslash(S_k \bigcup f_i)} sim(f_i, f_j)]^{\frac{1}{|V\backslash(S_k \bigcup f_i)|}} \tag{3.13}$$

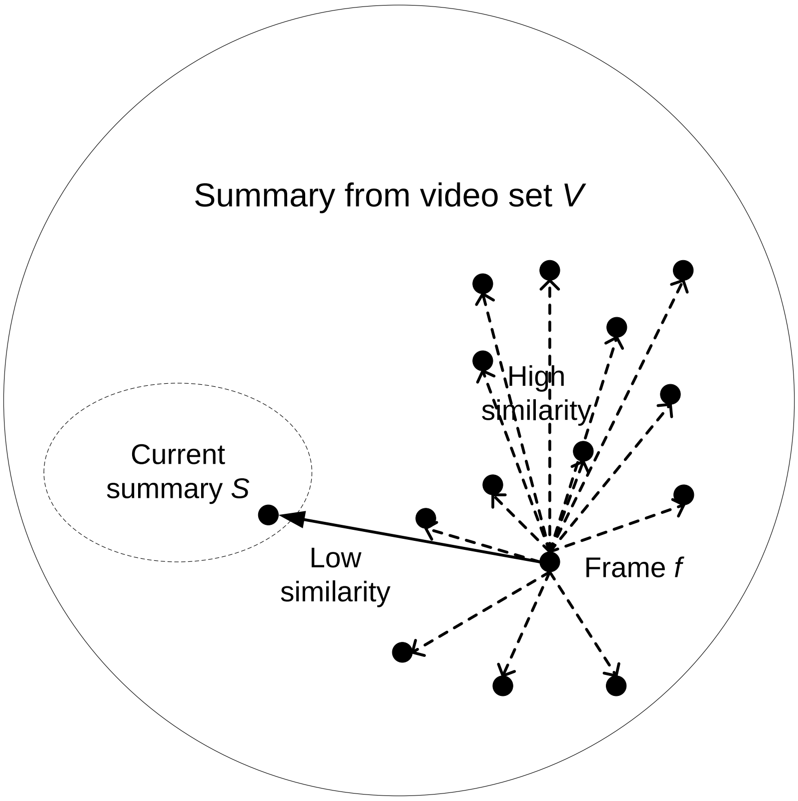


Figure 3.6: The illustration of Video-MMR

This leads to two variants: AM-Video-MMR and GM-Video-MMR. Both variants intend to model the amount of information that a new frame brings from the set of non-selected frames. Based on Video-MMR definition, the procedure of Video-MMR summarization is described as the following steps:

1. The initial video summary S_1 is initialized with one frame f_1, defined as:

$$f_1 = \arg\max_{f_i} \prod_{j=1, f_i \neq f_j}^{n} Sim(f_i, f_j)^{1/n} \tag{3.14}$$

   where f_i and f_j are frames from the set V of all frames from all videos, and n is the total number of frames except f_i, $n = |V| - 1$.

2. Select the frame f_k by Video-MMR:

$$f_{k+1} = \arg\max_{f_i \in V \backslash S_k}(\lambda Sim_1(f_i, V \backslash S_k) - (1 - \lambda) \max_{g \in S_k} Sim_2(f_i, g)) \tag{3.15}$$

3. Set $S_{k+1} = S_k \bigcup \{f_{k+1}\}$.

4. Iterate to step 2 until S has reached the desired size.

This algorithm has two variants, depending on which variant of the Video-MMR formula is used. Another issue is the value of the parameter λ, which can be used to adjust the relative importance of relevance and novelty. The next question is to select the best variant and value of λ. This will be explained in the next section.

3.2.3 Experimental results of Video-MMR

3.2.3.1 Experimental video sets

For our experiments, we have used two different video corpora from ”Wikio.fr”, which are not only used in this chapter, but also in the future chapters:

1. A small scale video corpus includes two groups ”DATI” and ”YSL”. This corpus is part of other experiments that are conducted on the text items associated in the groups. For the video part, ”DATI” includes 16 videos, while ”YSL” has 14 videos. The ”DATI” set contains videos about a French politician woman: most are directly captured from TV news, showing either the person herself, or people commenting her actions. The ”YSL” set contains videos related to the death of a famous designer. Some videos represent the burial, some are interviews or comments, some replay older fashions shows.

2. A large scale corpus of 89 sets of videos, each set containing videos collected from various sources, but dealing with the same event. Every set includes between 3 and 15 individual videos, for a total of more than 500 videos. Some videos are almost duplicates, for example the same video which has been published by different sources; some videos are quite different: one might show the actual event itself while another shows a comment about it.

3.2.3.2 The weight λ in Video-MMR

In the first step, we use the large video corpus containing 89 sets of videos from ”Wikio.fr”, and we try to find which combination (method and parameter values) provides the best results. We need to mention that in this thesis we subsample the video sequence by extracting one keyframe every 25 frames, so that one keyframe represents the visual content of one second sequence.

In the following experiments, the similarity of two frames, $sim(f_i, f_j)$, is defined as cosine similarity of visual word histograms:

$$sim(f_i, f_j) = cos(H_{f_i}, H_{f_j}) = \frac{H_{f_i} \cdot H_{f_j}}{\parallel H_{f_i} \parallel\parallel H_{f_j} \parallel} \tag{3.16}$$

where H_{f_i} and H_{f_j} are histogram vectors of frame f_i and f_j. To define visual words, we first detect Local Interest Points (LIPs) in the image, based on the Difference of Gaussian and Laplacian of Gaussian, then compute a SIFT descriptor. The SIFT descriptors are clustered into 500 groups by K-means to compose a visual vocabulary with 500 words. In [Lowe, 1999] the author uses 1000 keys, and in [Lowe, 2004] uses 536 words to represent a figure. It is obvious that more words in BOW can produce better experimental results, but we finally choose the number of words as 500 to reduce the computation redundancy. And the in a video the important frames with more words and unimportant frames with less words both exist, so a small number of words is able to compromise the quality and computation time. The processing software to get visual word histogram is Local Interest Point Extraction Toolkit (LIP-VIREO) [Video Retrieval Group, City U. of Hong Kong].

For a given set of videos, we have to compare two possible variants of the algorithm and find the best possible value of the parameter λ. To select the best combination, we use the criterion of the minimum $d(S, V)$ according to Eq. 3.8. The minimum distance displays the most similar summary with original video set V, and then the parameters belonging to this

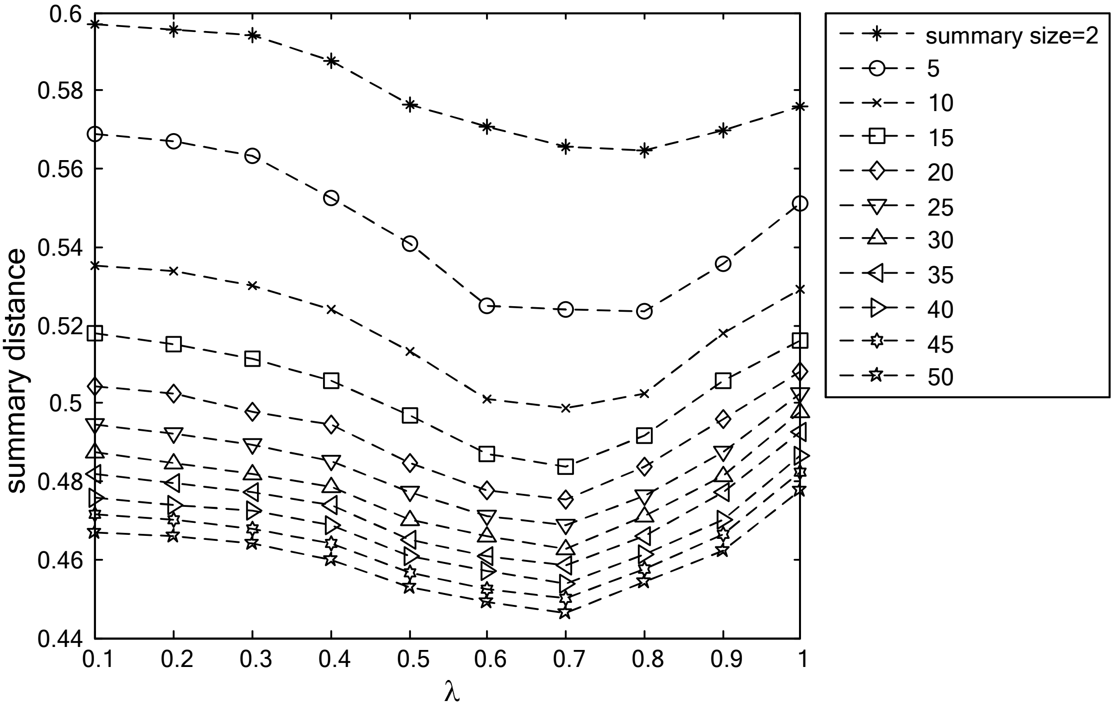


Figure 3.7: SRC of AM-Video-MMR

distance are the required parameters for the best Video-MMR. By sampling the possible values of λ into 0.1, 0.2, 0.3, ..., 0.9, 1.0, we obtain a total of $2 \times 10 = 20$ combinations. This method is named as Summary Reference Comparison (SRC). Fig. 3.7 shows SRC of AM-Video-MMR, whose summary size varying from 2 to 50 frames. Summary distances in Fig. 3.7 are the average distances of 89 videos sets. Fig. 3.8 shows the same summary distances of GM-Video-MMR.

From Fig. 3.7 and Fig. 3.8, we could conclude that the average distance is globally the minimum, when $\lambda = 0.7$ and the variant is AM-Video-MMR. So this combination is the best for Video-MMR. And in the following experiments and the following chapters of this thesis, we would use these values of the parameters.

3.2.3.3 Global and individual summarization

An easy way of generating a multi-video summary is to independently summarize each individual video in the video set and concatenate these summaries into a single one. This process is fast and easy to implement, but it ignores the inter-relations among different videos, so that similar keyframes could be selected in different individual summaries. We call this type of multi-video summarization as Individual Summarization (IS). In the algorithm that we have proposed, all videos are considered together, and keyframes are selected globally, so we call this process as Global Summarization (GS). Because Global Summarization considers both inter- and intra- relations of individual videos simultaneously, it should avoid the redundancy of Individual Summarization.

We retain the variant AM-Video-MMR and $\lambda = 0.7$ for the remaining experiments. We now construct the summary for a set of videos with two methods:

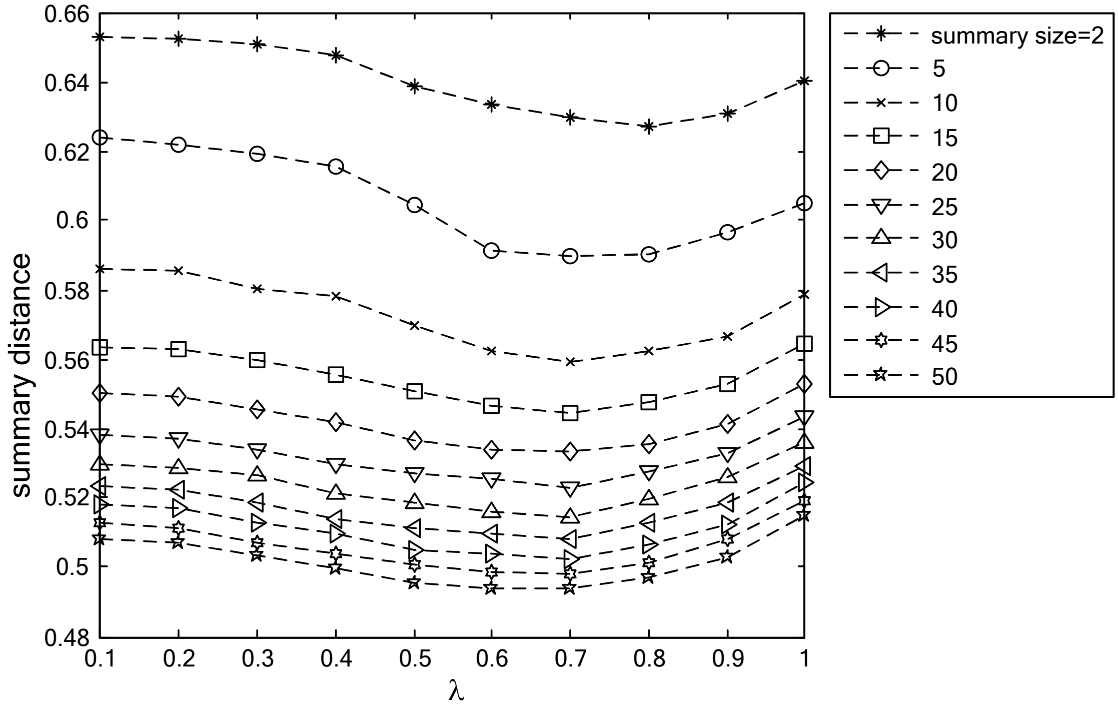


Figure 3.8: SRC of GM-Video-MMR

- GS (Global Summarization) as previously described;
- IS (Individual Summarization) by constructing a summary for each video in the set, and concatenating those summaries (no removal of possible duplicates).

We evaluate those summaries by computing their distances as Eq. 3.8 to the set of videos. We repeat experiments for different summary sizes. Fig. 3.9 shows an example of distances evolution for video set "YSL" whose summary sizes range from 1% to 15% of the size of the original video set. When we compare GS and IS for the summaries with the same size, GS distance is substantially lower than the IS distance, so Global Summarization is preferable to the Individual Summarization. Following experiments use GS. We repeated the experiment for all 89 video sets with summary percentage, 1%, 2%, 5% and 10% of original video. Most IS has larger summary distance than GS, and more than 85% of distance differences between IS and GS are less than 0.15.

3.2.3.4 Comparison with user-made summary

Human evaluation is commonly considered as ground truth. So it is meaningful to compare Video-MMR to human choice. In DATI and YSL video sets, 6 videos with the most obvious features were chosen. Then to obtain user-made summaries, we requested each of 12 people to select the 10 most important keyframes from all shot keyframes of those 6 videos. For the selected keyframes, the number of times they have been selected by a user is considered as a weight w. For example, if the number of selection of a keyframe is 3, then $w = 3$. A keyframe that has never been selected by any user has a weight of 0. Similar to Eq. 3.8,

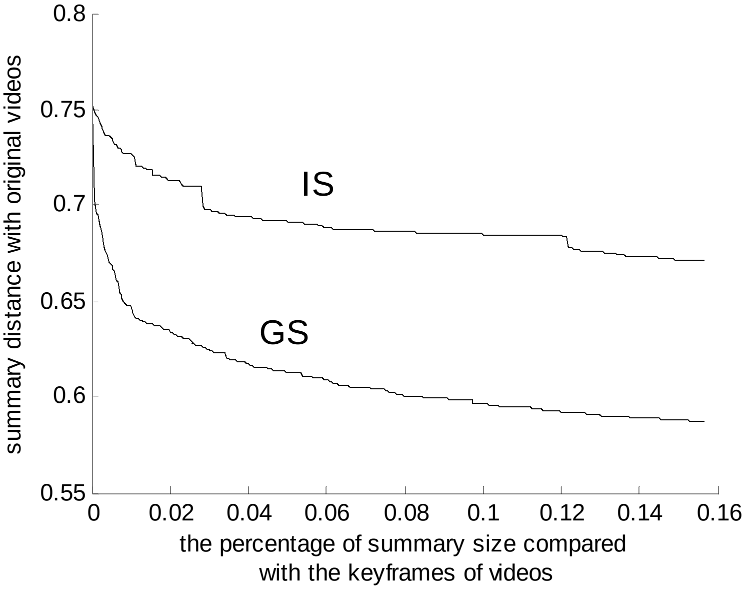


Figure 3.9: Comparison of "GS" and "IS" of video set "YSL"

the summary quality of Video-MMR with respect to the human choice can be defined as:

$$QC_{\textit{Video-MMR}} = \frac{1}{m}\sum_{i=1}^{m} w_i \cdot \max_{f \in S} sim(f, g_i) \tag{3.17}$$

where m is the number of keyframes of the video set, and f is a frame of Video-MMR summary, S. For further comparison, we also introduce the mean quality of every user-made summary compared with the other 11 user-made summaries:

$$QC_{human} = \frac{1}{N}\sum_{n=1}^{N} QC_n \tag{3.18}$$

where $QC_n = \frac{1}{m'}\sum_{i=1}^{m'} w_i \cdot \max_{f \in S_n} sim(f, g_i)$. In Eq. 3.17, $N = 12$, and m' is the unique keyframes' size of the other 11 user summaries, and frame f belongs to summary S_n. In this way, we can compare summary quality between Video-MMR, K-means and human choice (at least for a summary size of 10 keyframes).

$QC_{\textit{Video-MMR}}$ of Video-MMR summaries of "YSL" with 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 frames are shown in Fig. 3.10. We could see that $QC_{\textit{Video-MMR}}$ increases with the increase of summary size, because more similar information with human selection is included in Video-MMR summary. And $QC_{\textit{Video-MMR}}$ monotonically increases with summary size, which proves that the summary quality of Video-MMR is stable. Here we compare Video-MMR with K-means algorithm, which is a usual and popular clustering and summarization method. To construct the summary, K-means could select one frame from each cluster, but it cannot maximally promise the differences between the selected frames from the clusters because we need the particular and uncommon frames in video summaries for the video. Consequently the content diversity of the frames in K-means summary cannot be guaranteed. However, Video-MMR constructs the summary by selecting the frame which is the most different with the current summary. Therefore it is reasonable

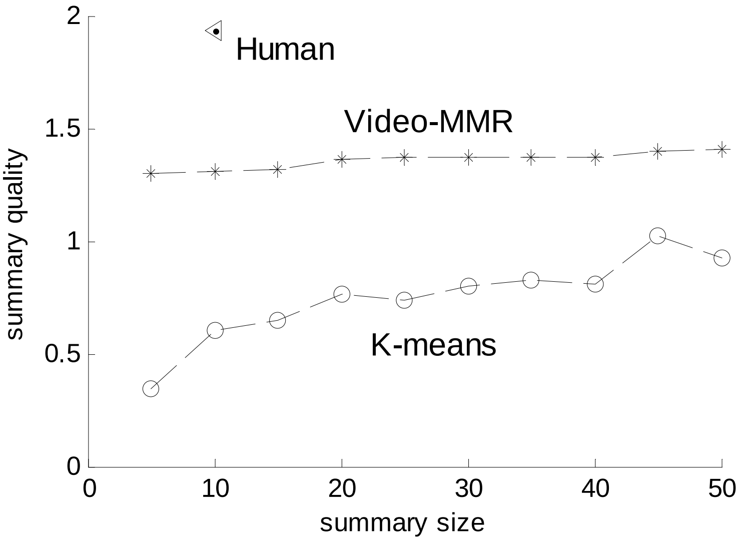


Figure 3.10: $QC_{\textit{Video-MMR}}$, $QC_{\textit{K-means}}$ and QC_{human}

to expect Video-MMR summary better than K-mean summary and it can represent more video content. Here summary quality of K-means is defined as $QC_{\textit{K-means}}$ like Eq. 3.17 too. $QC_{\textit{K-means}}$ evaluated by user-made summary is also shown in Fig. 3.10. It illustrates that summary qualities of Video-MMR are better than those of K-means, because the values of $QC_{\textit{K-means}}$ are smaller than $QC_{\textit{Video-MMR}}$. $QC_{\textit{K-means}}$ is not monotone, so summary quality of K-means is not as stable as Video-MMR. This is mainly caused by the random property of initial centers of K-means algorithm. Compared with K-means, Video-MMR has some advantages:

1. It avoids random initial centers, so it is more stable than K-means algorithm,
2. It considers inter- and intra- relations of summary,
3. It could control summary size because of the property of the incremental algorithm, but K-means is hard to control the final summary size,
4. It achieves dynamical summarization, which means that it could compute larger summary based on existent summary with smaller size.

Furthermore, QC_{human} is shown in Fig. 3.10. $QC_{\textit{Video-MMR}}$ with size 10 is closer to this ground truth than that of K-means.

Finally, AM-Video-MMR experimentally shows to be better in two variants, so Video-MMR in the following chapters refers to AM-Video-MMR. Experiments also show that Global Summarization is preferable to Individual Summarization. At last, the summary quality of Video-MMR is evaluated by human results as ground truth. We also demonstrate that not only the summary quality of Video-MMR is better than K-means, but also Video-MMR owns several advantages compared with K-means.

3.3 Conclusion of Video-MMR

First we have used several visualization approaches to present the relationships between keyframes of different videos in a video set. Then we have proposed Video-MMR by borrowing the idea from MMR and demonstrated that Video-MMR outperforms K-means and is closer to human assessment. Furthermore, GS has been proved better than IS for the multi-video summarization. A limit of Video-MMR is that it only exploits the low-level visual feature of the frames, but does not consider the temporal and semantic features in videos. Furthermore, Video-MMR is a generic algorithm for different genres of videos, but it is possible that different weights benefit different genres of videos.

Chapter 4

Video-MMR2

The more efficient approach by only exploiting the visual information is always important for the video summarization, because visual information is the most significant and obvious information in three kinds, text, visual and audio information. Therefore, we will discuss more about our algorithm, Video-MMR, and try to improve it after the discussion of its limits.

4.1 The discussion about Video-MMR

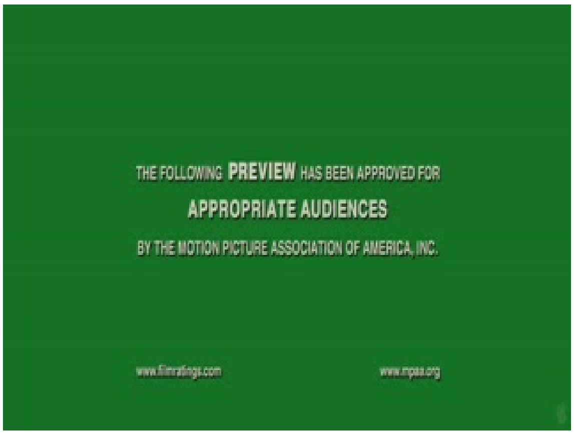
During the experiments of Video-MMR, we find it cannot work effectively when several special situations happen:

1. **The frame with a major color**

   When a big part of frames are of the same color, like the green color in the beginning and ending frames of many movie trailers and the dark color in the beginning frames of a news video. They are shown in Fig. 4.1 and Fig. 4.2.

   The frames like Fig. 4.1 and Fig. 4.2 can be selected during the procedure of Video-MMR as shown in Fig. 4.3, but they are visually unimportant and disliked by the user, though their text words burnt in the frame are important in another sense.

2. **The dull frame**



THE FOLLOWING **PREVIEW** HAS BEEN APPROVED FOR

APPROPRIATE AUDIENCES

BY THE MOTION PICTURE ASSOCIATION OF AMERICA, INC.

www.filmratings.com www.mpaa.org

Figure 4.1: The frames with a major color in green color

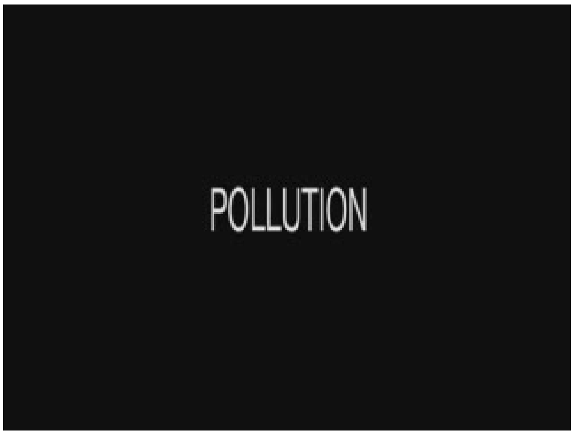


Figure 4.2: The frames with a major color in black color

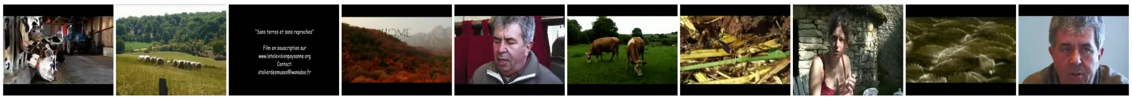


Figure 4.3: A summary with a frame with a major color in black

In some frames, the content in the frame is interesting for the user, but it is too dull to discern the detail in it. An example of dull frames is shown in Fig. 4.4, which is a frame in the summary shown in Fig. 4.5. It is hard for the user to review the content represented in this kind of frames. Therefore, it is better to avoid selecting these frames in the summary.

3. **The duplicated frame**

   The last and most important limit of Video-MMR is that it may select duplicated frames when there are many duplicated frames in the original video. The value of Sim_1 of Video-MMR formula, Eq. 3.11, for a frame increases with the increase of the number of duplicated frames or very similar frames for this frame, so Video-MMR favors the frames with many duplicates or very similar frames in the video.

Figure 4.4: A dull frame

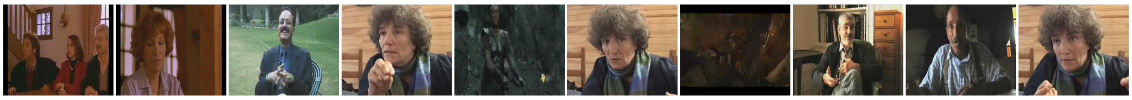


Figure 4.5: A summary with a dull frame

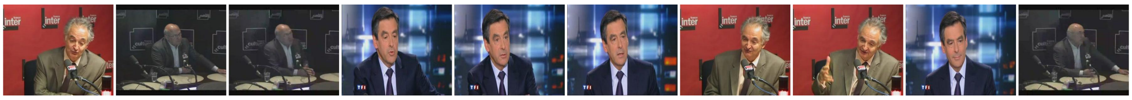


Figure 4.6: A summary with the duplicates

From Fig. 4.5 we can see that there are similar frames with the face of the same person in the summary. And we display one extreme example in Fig. 4.6, which is created by Video-MMR from the original frames shown in Fig. 4.7.

For the summary shown in Fig 4.6, it is reasonable to include some duplicates for a summary with size of 10, because the individual frames in Fig. 4.7 are 6. But the problems are:

- The duplicates are selected first, not the fresh frames for the summary in the incremental mechanism of Video-MMR.
- It also causes that some fresh frames for the summary in Fig. 4.7 are not selected in the summary.

Totally, we need to improve Video-MMR in these 3 aspects, especially the duplicates. We will propose the improved algorithm in the following section.

4.2 The principle of Video-MMR2

4.2.1 Exclude visually unimportant frames

1. **Exclude the frames with a major color**

   To exclude the frames with a major color, we simply use the color histogram. In the histogram of RGB model, we consider 64 bins for each channel, then totally we got $64 * 64 * 64 = 262144$ bins for RGB, stored in the form of one-dimension. The reason to use a large number of bins is that in most situations the major color in a excluded frame is only one color, not multiple colors, as shown in Section 4.1. Then

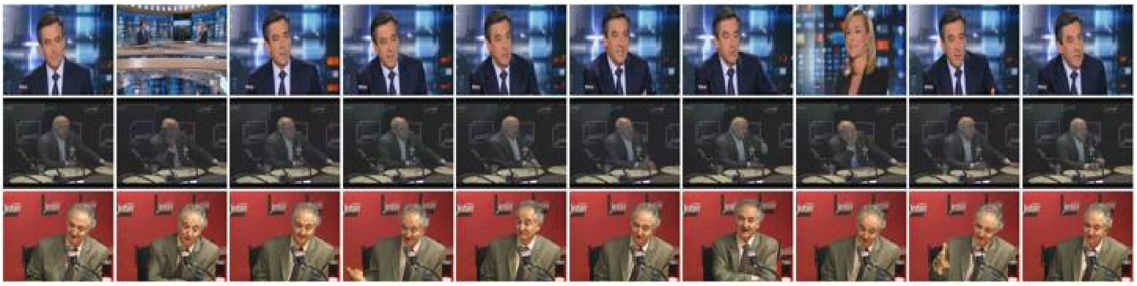


Figure 4.7: The original frames for Fig. 4.6

for the continuous bins in a frame, if they satisfy the following criterion, Eq. 4.1, this frame is excluded from the possible candidates for the summary:

$$\sum_{i=j-1, j \in M}^{j+1} H(i) > \varepsilon \tag{4.1}$$

where H is the RGB histogram with 262144 bins, $M = \{2, \ldots, 262143\}$, and ε is an empirical value, 0.55.

2. **Exclude the dull frames**

   To remove the dull frames from the candidate frames for the summary, we decide to simply exclude the frames with extreme low value for the Value channel in HSV histogram. By mimicking the method of excluding the frames with a major color, we first compute HSV histogram of a frame with 16 bins for each channel. Then the frames with the criterion described in Eq. 4.2 are excluded from the set of the candidate frames.

   $$\nu(i) > \delta \tag{4.2}$$

   where $i = 1$ in the bins $\{1, \ldots, 16\}$ and ν is V channel of HSV histogram. $\delta = 0.55$ according to our experiments.

4.2.2 Remove the duplicates

In Section 4.2.1 we have described how to remove visually unimportant frames, but these frames only occupy a little part in the summary. The most important factor decreasing the quality of the summary is the duplicated frame. The approach to avoid the duplicate is to reduce the importance of the frame with a large amount of same or similar frames, which means decreasing Sim_1 in Video-MMR formula even there are duplicates for a frame. Consequently, we propose a novel formula, Video-MMR2, in Eq. 4.3.

$$S_{k+1} = S_k \bigcup \arg\max_{f_i \in V \backslash (S_k \bigcup f_i)} (\lambda Sim_1(f_i, V \backslash (S_k \bigcup f_i)) - (1-\lambda) \max_{g \in S_k} Sim_2(f_i, g)) \tag{4.3}$$

where

$$Sim_1(f_i, V \backslash (S_k \bigcup f_i)) = \frac{\sum_{f_j \in V \backslash (S_k \bigcup f_i)} sim(f_i, f_j)(1 - \max_{g \in S_k} sim(f_j, g))}{\sum_{f_j \in V \backslash (S_k \bigcup f_i)} (1 - \max_{g \in S_k} sim(f_j, g))} \tag{4.4}$$

and $Sim_2(f_i, g)$ is the same as Video-MMR.

4.2.3 The procedure of Video-MMR2

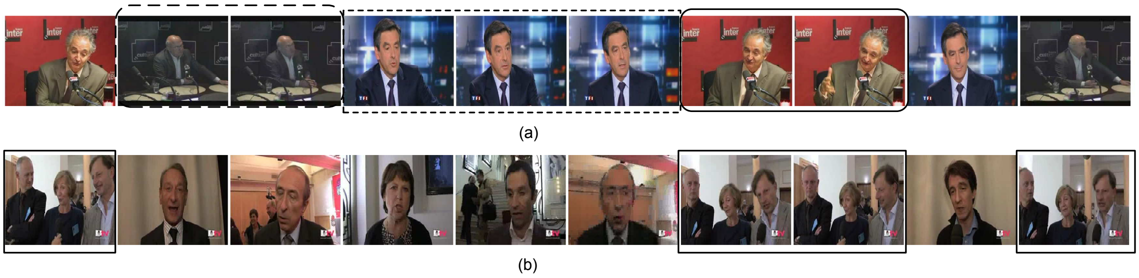
The procedure of Video-MMR2 summarization is described as the following steps:

1. Exclude the visually unimportant frames from V as described in Section 4.2.1.

2. The initial video summary S_1 is initialized with one frame f_1, defined as:

   $$f_1 = \arg\max_{f_i, f_i \neq f_j} (\prod_{j=1}^{n} Sim(f_i, f_j))^{1/n} \tag{4.5}$$

   where f_i and f_j are frames from the set V of all frames from all videos, and n is the total number of frames except f_i.



(a)

(b)

Figure 4.8: The hard frames (in box) for visual word histogram: Similar/same frames are in same kind of box

3. Select the frame f_k by Video-MMR2, Eq. 4.3.
4. Set $S_k = S_{k-1} \bigcup \{f_k\}$.
5. Iterate to step 3 until S has reached the desired size.

4.3 Experimental results of Video-MMR2

We want to check if the visually unimportant frames and the duplicated frames are removed from the summary compared with Video-MMR, and if Video-MMR2 could still produce good summaries.

If the original frames of original videos are too many, it is hard for human to make a decision to above questions. Therefore, we decide to preprocess the frames from original videos as the approach in VERT explained in detail in Section 6.4. Then the candidate frames for Video-MMR2 are 30-60 frames from 3-6 videos for a video set. Totally we have 28 video sets whose genres are *News*, *Movie* and *Documentary*.

In addition, the similarity used in Video-MMR2 is different from the one in Video-MMR. During our experiments, we find that the similarity of visual word histogram is weak to represent the similarity of the frames with the same or similar object in the same or similar background, especially the face which appears for a lot of times in our experimental videos. In Fig. 4.8 we demonstrate two summaries containing the frames hardly processed by Video-MMR with visual word histogram.

Therefore, for Video-MMR2 we use the similarity between two frames as Eq. 4.6.

$$sim(f_i, f_j) = (sim_{visual\ word}(f_i, f_j) + sim_{color}(f_i, f_j))/2 \tag{4.6}$$

where $sim_{visual\ word}(f_i, f_j)$ means the similarity of visual word histogram between the frame f_i and f_j, and $sim_{color}(f_i, f_j)$ means the similarity of color histogram between the frame f_i and f_j. After adding the similarity of color histogram, it is easier for Video-MMR2 to discern the moving objects in the frame, and summaries by Video-MMR2 are shown in Section 4.3.2.

4.3.1 SRC2

In Section 3.2.3 we use SRC to decide the weight λ in Video-MMR. For Video-MMR2 we face up to the same problem of choosing a optimized weight λ. However, Video-MMR2 tries to avoid the duplicates and cover more visual content represented in the video, which

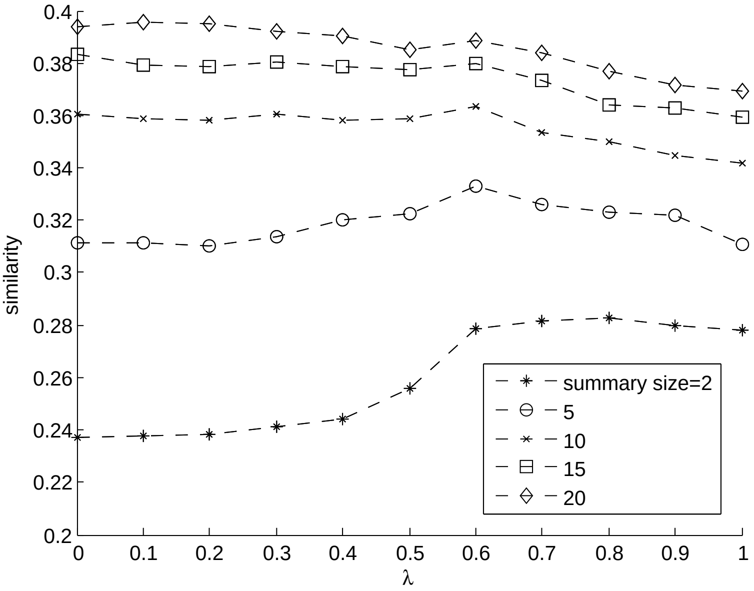


Figure 4.9: SRC2

does not promise the maximum similarity with videos if many duplicated frames exist, so we cannot directly exploit SRC by comparing the similarity of visual word histogram between the summary and videos.

Therefore, we propose a novel formula to define the distance or similarity between the summary and the original videos in Eq. 4.7. In Eq. 4.7 we compute the visual coverage represented by the summary:

$$Sim_{src2} = \frac{1}{N}\sum_{i=1}^{N}[1 - \max_{f\in R_i(V\backslash S)}(\min_{g\in S} d(f,g))] \tag{4.7}$$

where $R_i(V\backslash S)$ is a random selection of k frames from $V\backslash S$, and in our experiment we set $k = 20$ and $N = 20$ because the frame number in a video set is from 30 to 60 in our experimental data.

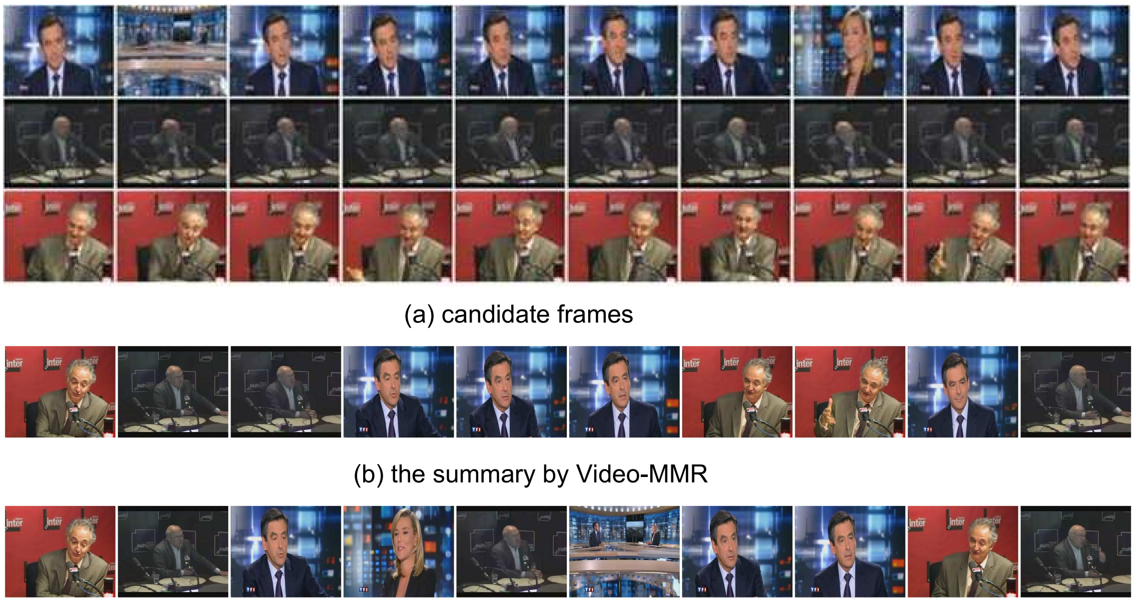
The values Sim_{src2} of $\lambda = 0.0, 0.1, \ldots, 1.0$ when the summary size is 2, 5, 10, 15 or 20 are shown in Fig. 4.9.

From Fig. 4.9 we can find that $\lambda = 0.6$ commonly achieves the maximal visual coverage in the summary compared to candidate frames for different summary sizes. So it is reasonable to set λ as 0.6 in Video-MMR2 formula.

4.3.2 The improvement of the summary

Fig. 4.10, Fig. 4.11 and Fig. 4.12 show three examples of original candidate frames, their summaries by Video-MMR and their summaries by Video-MMR2. In Fig. 4.10(a), Fig. 4.11(a) and Fig. 4.12(a) frames in one row are from the same video.

- Fig. 4.10(c) shows that Video-MMR2 selects the fresh frames first except 5th frame which is caused by the optimized λ for all the video sets not only for this video set. But video-MMR2 still selects all the individual frames from 4.10(a) not like 4.10(b). The other duplicates in 4.10(c) are impossible to avoid because the number of individual frames is limited in the original frame, 4.10(a).



(a) candidate frames

(b) the summary by Video-MMR

(c) the summary by Video-MMR2

Figure 4.10: First example of Video-MMR2

- Fig. 4.11(b) includes a black frame, 3th frame, which is excluded from Fig. 4.11(c). And 10th frame in Fig. 4.11(b), a face also appearing in 5th frame in Fig. 4.11(b) is removed by Video-MMR2 in Fig. 4.11(c).

- In Fig. 4.12(b) the summary by Video-MMR is good compared to candidate frames in Fig. 4.12(a). So the modification in Fig. 4.12(c) is little, and most frames are kept. And the summary by Video-MMR2 in Fig. 4.12(c) is still a good summary.

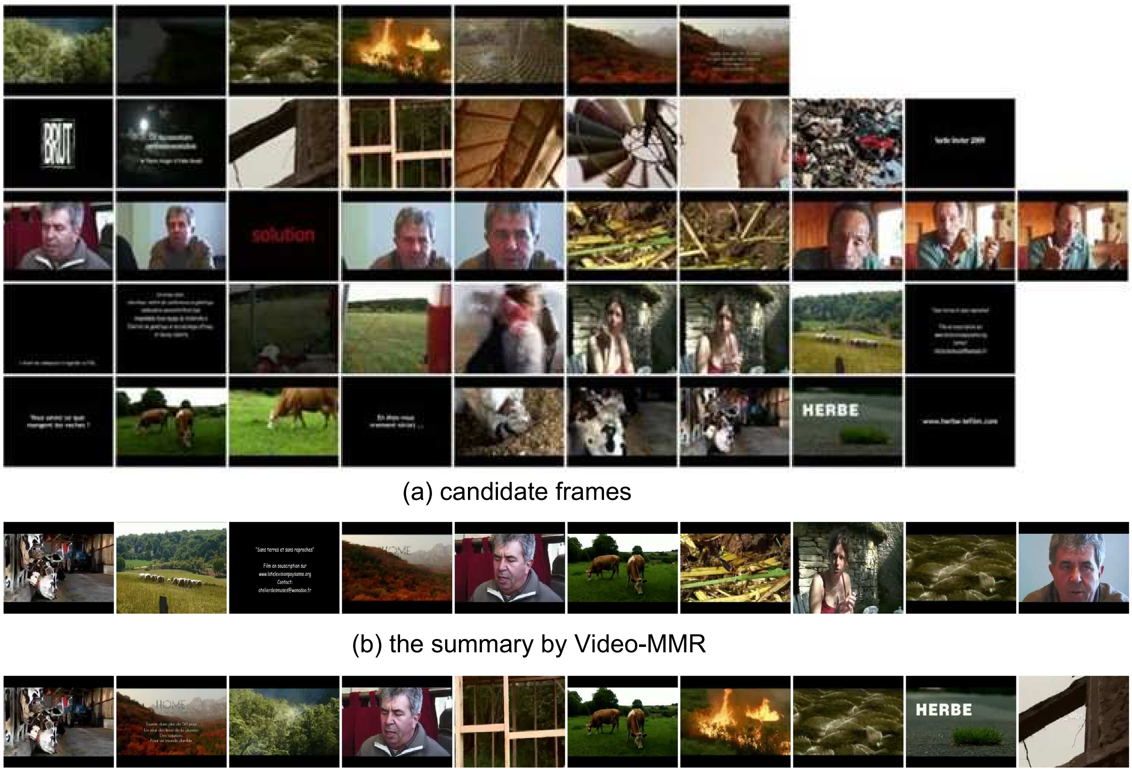
From the analysis of Fig. 4.10, Fig. 4.11 and Fig. 4.12, we can see that the performance of Video-MMR2 is better than Video-MMR. Furthermore, for 28 video sets in Table 4.1 we display the suppressed redundant frames by Video-MMR2 compared to the summaries of Video-MMR, the remaining redundant frames in summaries by Video-MMR2 and the possible improvement for redundant frames in summaries by Video-MMR2 (here the summary size is 10).

From Table 4.1 we can see that many redundant frames in summaries by Video-MMR are removed by Video-MMR2, and the remaining redundant frames in summaries by Video-MMR are impossible to be removed because the videos are not sufficiently diverse to provide 10 different non-similar keyframes. Furthermore, we can compute the average performance of Video-MMR2 from Table 4.1:

1. Average number of suppressed redundant frames is 0.5 frames.

2. Average number of remaining redundant frames is 0.39 frames, including those which are impossible to remove.

3. Average number of unique frames in summaries by Video-MMR2 is 9.61 frames.

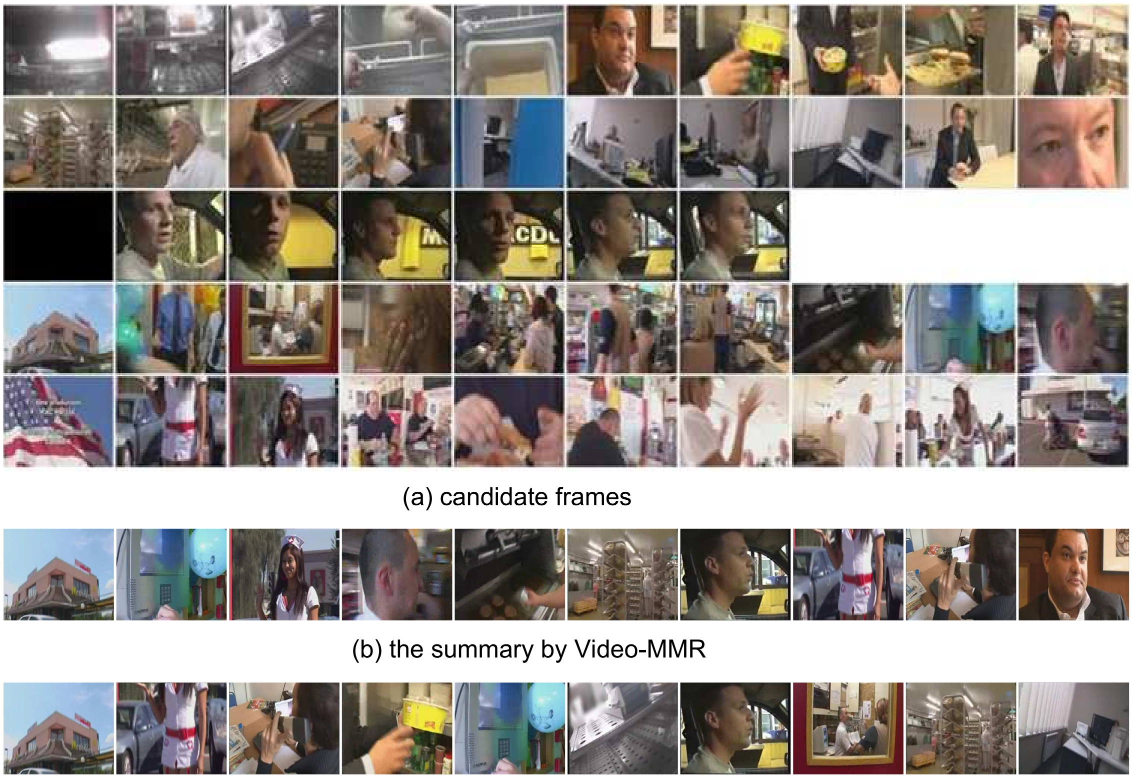


(a) candidate frames

(b) the summary by Video-MMR

(c) the summary by Video-MMR2

Figure 4.11: Second example of Video-MMR2



(a) candidate frames

(b) the summary by Video-MMR

(c) the summary by Video-MMR2

Figure 4.12: Third example of Video-MMR2

Table 4.1: The improvements brought by Video-MMR2

Video name	Suppressed Redundant frames	Remaining redundant frames	The possible improvement of redundancy
1269902	0	0	0
26916	0	0	0
2923014	3	0	0
299483	0	0	0
2997420	0	0	0
3010847	1	0	0
3042273	0	0	0
3012751	0	0	0
3194299	1	5	0
354610	1	0	0
390539	1	2	0
459166	1	4	0
504701	0	0	0
534553	0	0	0
548083	0	0	0
590303	1	0	0
658217	2	0	0
665394	0	0	0
678269	0	0	0
687626	0	0	0
692301	0	0	0
718919	0	0	0
745348	0	0	0
759309	0	0	0
762796	0	0	0
780944	0	0	0
781047	1	0	0
841675	2	0	0

4. We find that average number of unique frames in ideal summaries by human is 9.61 frames. Therefore, average number of possible improvement compared with ideal summaries is 0 frames.

4.4 Conclusion of Video-MMR2

In this chapter, we have proposed Video-MMR2 to improve the previous summarization algorithm, Video-MMR, and reform 3 aspects of the limit in Video-MMR. The summarization procedure of Video-MMR2 avoids selecting the frame with a major color, the dull frame and the duplicated frame, which have been proved by the experimental results. Especially for the duplicated frame appearing in the summary of Video-MMR, our experimental results have demonstrated that the summary by Video-MMR2 avoids this limit and achieves the best. Therefore, Video-MMR2 keeps the advantages of Video-MMR and avoids the limits of Video-MMR, so Video-MMR2 is a good improvement on Video-MMR when we only exploit visual information of the video.

Chapter 5

Multimedia MMR

In Chapter 3, we have presented the principle of Video-MMR. Then we proposed Video-MMR2 in Chapter 4 by using visual information too. However, Video-MMR and Video-MMR2 only exploits the visual information in the video, and the audio and text information in the video is discarded, which is obviously significant for capturing the important content of the video. Consequently, we extend Video-MMR and propose several algorithms to summarize multi-video by exploiting audio and text information of the video too.

In Section 5.1, we simply add an audio part into the formula of Video-MMR by mimicking the formula of Video-MMR, and propose AV-MMR (Audio Video MMR). However, the simple extension of Video-MMR with the audio feature misses the relation of visual and audio information at the semantic level. Balanced AV-MMR (Balanced Audio Video MMR) in Section 5.2 eliminates this limit by considering the relation of visual and audio information and the temporal information of the multi-video. In Balanced AV-MMR various parameters are manually decided, so we decide to propose an algorithm with the automatic optimization of the parameters, OB-MMR (Optimized Balanced Audio Video MMR).

In addition to exploit audio information, we also propose an algorithm, TV-MMR (Text Video MMR) in Section 5.4, which uses both text and visual information by adding textual formula part into Video-MMR formula. Since TV-MMR could combine text and visual information at the same time, we propose a model of dynamic summary in Section 5.5.2 to optimally decide the best duration of segments on average in a summary with the form of video skimming. Similarly, the model of static summary based on Video-MMR is suggested in Section 5.5.1 to optimize numbers of text utterances and video frames in a predefined space containing both text utterances and stationary video frames.

5.1 Audio Video MMR

In the current techniques of video summarization, some algorithms are for a single video sequence [Yahiaoui et al., 2001] and [Money and Agius, 2007], and some recent research work begins to deal with multi-video summarization, like Video-MMR proposed in Chapter 3. Most current algorithms of video summarization only exploit visual features in the video, because it is hard to combine different multimedia cues in the processing. However it is a limitation for many videos containing important acoustic or textual information.

Some algorithms have been proposed [Sugano et al., 2002] [Furini and Ghini, 2006] [Xu et al., 2005] for summarization using both audio and video information. But these methods are often domain-specific, for example focusing on music video clips. In [Sugano et al., 2002], the authors utilize motion features based on MPEG-7 and detect highlight by analyzing audio class and audio level. In [Furini and Ghini, 2006] M. Furini removes some silent segments from original videos after detecting the silences in the audio. In an approach for music video summarization [Xu et al., 2005], the authors detect the chorus in audio and the repeated shots in video track. The generic approach of video summarization is still an important issue.

In this section we bring the audio information into the Video-MMR algorithm and propose a novel algorithm, Audio Video Maximal Marginal Relevance (AV-MMR) [Li and Merialdo, 2010b]. AV-MMR uses both the visual feature (Bag of Visual Words [Video Retrieval Group, City U. of Hong Kong]) and the audio feature (Mel-frequency cepstral coefficients (MFCCs) [Davis and Mermelstein, 1980] [Zheng et al., 2001] [IRISA]). Several variants of AV-MMR are compared. Besides the AV-MMR algorithm, we also propose a visualization tool to illustrate and to compare the audio and the visual content of a set of video sequences.

5.1.1 Circle representation of video and audio keyframes

In Section 3.1.2 we have proposed a visualization method of circle representation to illustrate the relation of frames in multiple videos. Therefore, we also wish to illustrate the relations of video and audio keyframes in multiple videos first in this section to find if different audio and video frames are coherent according to this circle approach.

First, we need to describe the video sets used in the circle representation. Similar to the video used in previous chapters, video sets in this section are 63 videos containing both video and audio channels in video sets mentioned in Section 3.2.3.1. Each video set has a specific topic, like a film or a ceremony. Each set has usually from 3 to 8 videos, with a maximum of 13 videos. The durations of most videos are from 1 minute to 7 minutes. These videos sets represent topics from different genres, such as films, documents, music clips, sport to advertisement.

To represent the audio content of a one second audio segment, we use the common Mel-frequency Cepstral Coefficients (MFCCs) [Zheng et al., 2001]. The software to get MFCC vectors, SPro Toolkit, is from [IRISA]. According to selected parameters, SPro creates 100 MFCC vectors per second, with a set of 21 cepstrum coefficients in each MFCC vector. We average these 100 MFCC vectors in a second for 21 coefficients to obtain the audio feature vectors with 21 coefficients, S_{MFCC}. The similarity between two averaged MFCC vectors is computed and normalized as:

$$sim_A(a_i, a_j) = 1 - \frac{|a_i - a_j|}{\max_{a_m, a_n \in S_{MFCC}}(|a_m - a_n|)} \tag{5.1}$$

where a_i, a_j, a_m and a_n are averaged MFCC vectors. To combine audio and video similarity measures efficiently, we first normalize and rescale them according to the following equation:

$$X' = \frac{X - \mu}{\sigma} \tag{5.2}$$

where X is the initial value, and X' is the normalized value. μ and σ are the mean and standard deviation of the original values. The final visual and audio similarity measures are respectively called sim'_I and sim'_A.

Visualization of the content of a set of videos is a useful tool for understanding the possible relations between the various videos. In Section 3.1.2 we exploit a circle space to visualize the relations of video keyframes from multi-video sets. The position of video frames in a circle are computed by Eq. 3.1 and Eq. 3.3, and the relation of video frames is illustrated in Fig. 3.3. And we conclude that if a frame is more similar to a specific video, the position of this frame will be closer to this video. If the position of a frame is closer to the center of circle space, it means that this frame has an average similarity with all the videos.

Eq. 3.1 and Eq. 3.3 are suitable for both image and audio frames. For image information $sim(f,k)$ can be replaced by $sim'_I(f,k)$; while for audio information, $sim(f,k)$ is changed to $sim'_A(f,k)$.

Two examples of circle representation for two different video sets are shown in Fig. 5.1. The first set contains 4 videos and the second contains 3 videos. On the circle boundary, the "o" are the points representing the videos. Inside the circle, the "×" are the points representing the audio content of keyframes, and the "o" are the points representing the visual content of keyframes. Because of the number of frames, the "×" and the "o" may overlap with each other in the figure. We can see in Fig. 5.1(a) and 5.1(b) that audio frames and video frames have some similarity and are not totally independent with each other. This representation provides an easy visualization of the content of a video set.

5.1.2 Audio Video Maximal Marginal Relevance

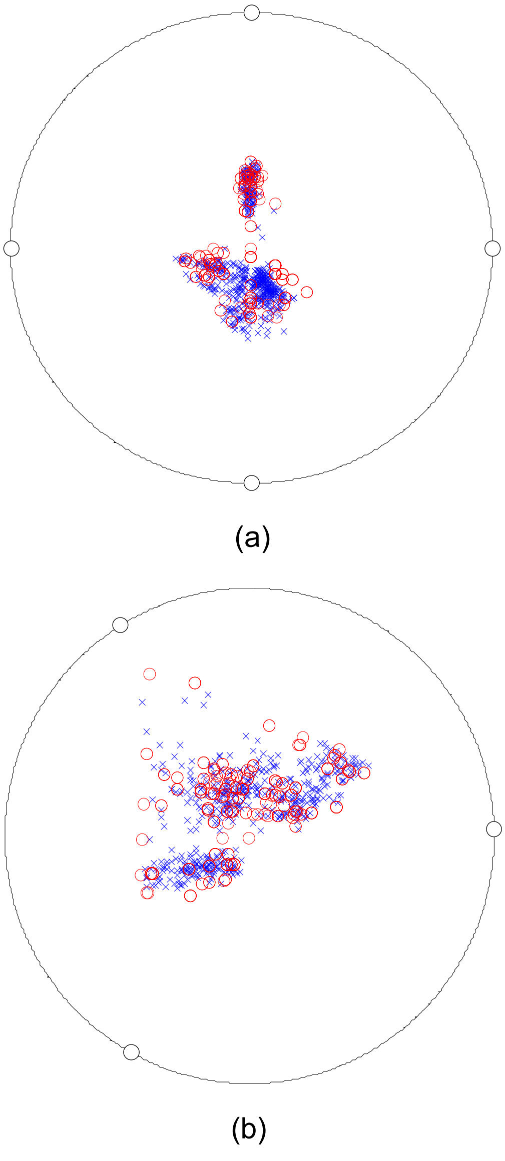
A video sequence contains both audio and video track. Here, we extend Video-MMR to Audio Video Maximal Marginal Relevance (AV-MMR) by considering information from both audio and video. To each keyframe, we associate the corresponding one second audio segment. We then exploit audio and visual features and modify Eq. 3.11 into Eq. 5.3, which defines how summary S_{k+1} can be constructed by iteratively selecting a new keyframe:

$$\begin{aligned} S_{k+1} = S_k \bigcup \operatorname*{arg\,max}_{f \in V \backslash S_k} [\lambda Sim_{I1}(f, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g) + \\ \mu Sim_{A1}(f, V \backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)] \end{aligned} \tag{5.3}$$

where Sim_{I1} and Sim_{I2} are the same measures as Sim_1 and Sim_2 in Eq. 3.11. Sim_{A1} and Sim_{A2} play roles similar to Sim_{I1} and Sim_{I2}, but use the audio information of f. Eq. 5.3 combines visual and audio similarities corresponding to the same frame, so we call this algorithm Synchronous AV-MMR (SAV-MMR). It is also possible to select audio and video independently, as in Eq. 5.4:

$$\begin{aligned} S_{k+1} = S_k \bigcup \operatorname*{arg\,max}_{f, f' \in V \backslash S_k} [\lambda Sim_{I1}(f, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g) + \\ \mu Sim_{A1}(f', V \backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f',g)] \end{aligned} \tag{5.4}$$

The algorithm based on Eq. 5.4 is named as Asynchronous AV-MMR (AAV-MMR). AAV-MMR removes the restriction that visual and audio content are selected from the same instant of the video sequence. AV-MMR also has two variants as Video-MMR: AM-AV-MMR and GM-AV-MMR. For AM-AV-MMR, the equations of Sim_{I1} and Sim_{A1} are the same as Eq. 3.12. GM-AV-MMR uses the definition in Eq. 3.13. Parameter λ controls



(a)

(b)

Figure 5.1: Circle representation: In the circle ”×”=audio frame;”o”=video frame. On the circle, ”o”=video.

the relative importance of relevance and novelty of selected visual information. Similarly, parameter μ plays the same role for audio information.

With these definitions of AV-MMR through Eq. 5.3 and Eq. 5.4, we can describe the complete AV-MMR summarization procedure as the following sequence of steps:

1. The initial video summary S_1 is initialized with one frame, defined as:

$$S_1 = \underset{f_i, f_i \neq f_j}{\arg\max}[\prod_{j=1}^{n} Sim_I(f_i, f_j) \cdot \prod_{j=1}^{n} Sim_A(f_i, f_j)]^{\frac{1}{n}} \tag{5.5}$$

where f_i and f_j are frames from the set V of all frames from all videos, and n is the total number of frames except f_i. Sim_I computes the similarity of image feature vectors between f_i and f_j; while Sim_A is the similarity of audio feature vectors between f_i and f_j.

2. Select the frame f_k by SAV-MMR or AAV-MMR. We only mention the SAV-MMR equation here:

$$\begin{aligned} S_{k+1} = S_k \bigcup \underset{f \in V \backslash S_k}{\arg\max}[&\lambda Sim_{I1}(f, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f, g) + \\ &\mu Sim_{A1}(f, V \backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f, g)] \end{aligned} \tag{5.6}$$

3. Set $S_k = S_{k-1} \bigcup \{f_k\}$

4. Iterate to step 2 until S has reached the predefined size.

In the following sections, we will search for the best values for the parameters λ and μ, and we will compare experimentally the two variants, AM-AV-MMR and GM-AV-MMR.

5.1.3 Experimental results of AV-MMR

5.1.3.1 Summary Reference Comparison

In Video-MMR of Chapter 3, we have compared the results of different variants and parameters to select the best variant and parameter values by Summary Reference Comparison (SRC). In this section, we use the same approach, SRC, to obtain the best variant and parameter values. For SRC, we first need to define the distance between a video summary and the original video. Since we consider both the visual and the audio information in AV-MMR, distance between a video set V and its summary S is defined as the following equation:

$$d(S, V) = \frac{1}{n} \sum_{j=1}^{n} \min_{f_j \in V, g \in S} [1 - \frac{sim'_I(f_j, g) + sim'_A(f_j, g)}{2}] \tag{5.7}$$

where n is the number of frames in the video, V. g and f_j are frames respectively from video summary, S, and V. Then the best summary (for a given length) is the one that achieves the minimum distance, and the values of the parameters which achieve the minimum summary distance will be kept as the best.

For a given set of videos, we have to compare two possible variants of the algorithm and to find the best possible values for the parameters λ and μ. For each parameter, we try the values $0.1, 0.2, 0.3, \ldots, 0.9$. This leads to a total of $9 \times 9 = 81$ combinations. Fig.

5.2 shows the evolution of the summary distance depending on the summary length. For simplicity, we only display the lines corresponding to the minimum distances for each audio parameter μ, because it is hard to display 81 curves. We only consider SAV-MMR, and we compare the two variants, AM-AV-MMR and GM-AV-MMR. In Fig. 5.2 the summary distance is the sum of audio summary distance and visual summary distance, so the value is possibly larger than 1.0.

Fig. 5.2 shows the SRC curves of AM-AV-MMR and GM-AV-MMR, for summary sizes varying from 5 to 50 frames, where each of the 9 curves corresponds to the minimum values obtained with parameter μ linearly ranging from 0.1 to 0.9. Summary distances in Fig. 5.2 are the mean distances over 62 video sets in 63 video sets mentioned in Section 5.1.1. The residual $63rd$ video set will be used to demonstrate the effect of AV-MMR later.

From these evaluations, we can see that the minimum summary distance is obtained in Fig. 5.2(a) with a value of $\mu = 0.5$. The corresponding value of the parameter $\lambda = 0.7$. The values of AM-AV-MMR are globally smaller than GM-AV-MMR, so we can conclude that AM-AV-MMR generates better summaries.

5.1.3.2 Comparison of Video-MMR and AV-MMR

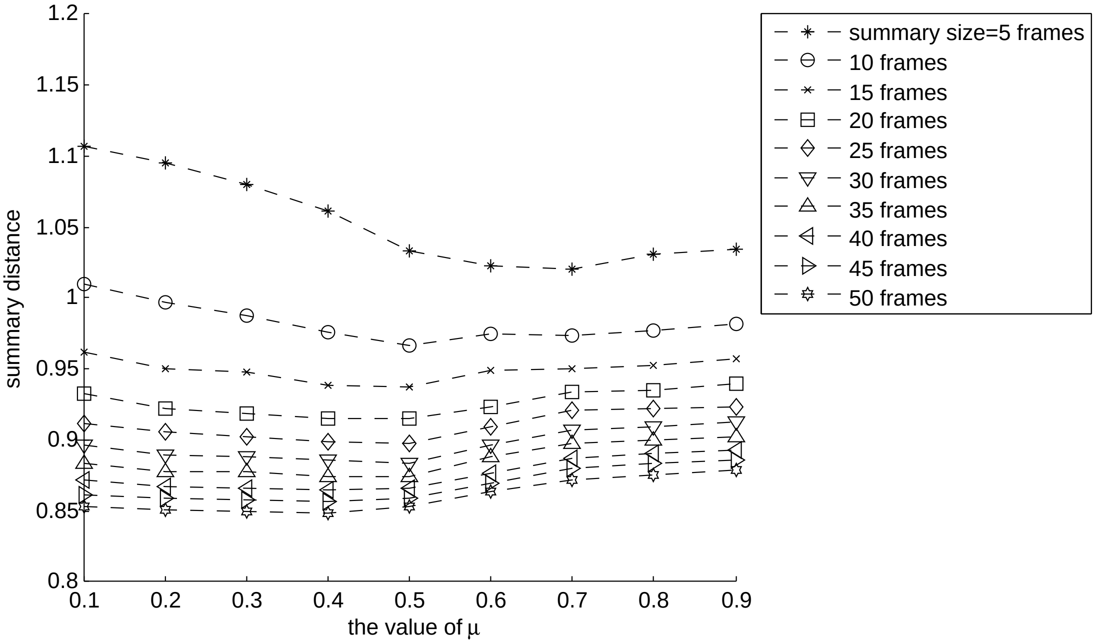
Once we have found the best values of the parameters, $\mu = 0.5$ and $\lambda = 0.7$, and the best variant of SAV-MMR, AM-AV-MMR, we want to compare the AV-MMR approach with the previous Video-MMR algorithm. We also want to compare the two synchronous and asynchronous variants of AV-MMR.

We use the 63rd video set, which has not been used during the training phase. This video set, whose name is ”YSL”, is the same as the ”YSL” video set mentioned in Chapter 3, containing 14 videos. We run all three summarization algorithms, Video-MMR, SAV-MMR and AAV-MMR, to generate summaries with sizes from 1 to 50 keyframes. Then, we compute the summary distance for each generated summary and display the results in Fig. 5.3.

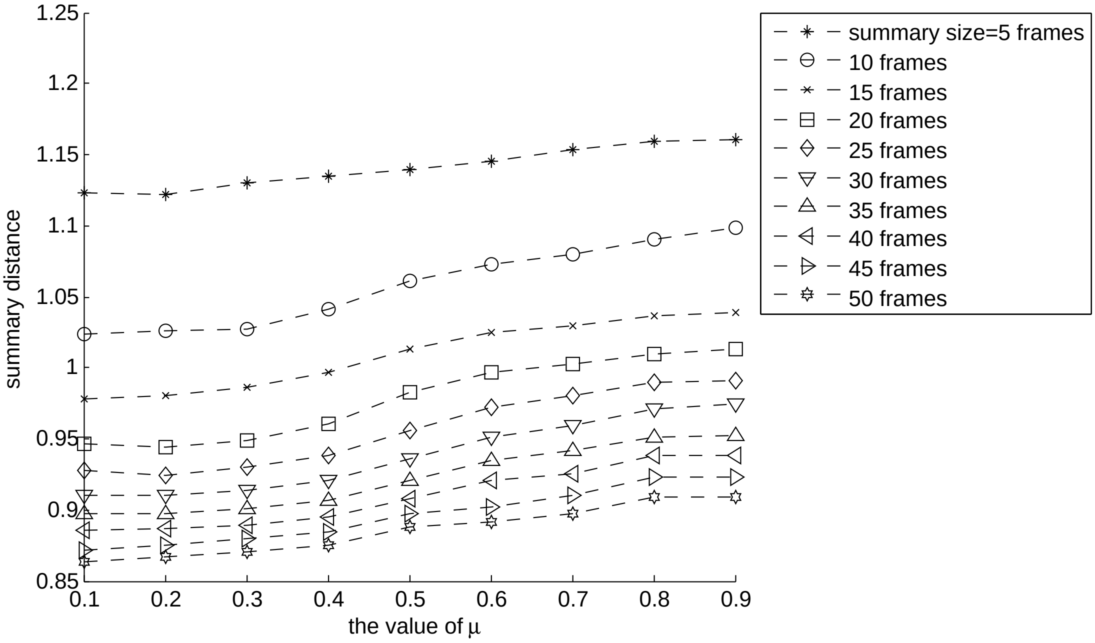
From Fig. 5.3, it is clear that AV-MMR is better than Video-MMR. This is natural, since Video-MMR does not consider the audio information from the original sequence, so it is not able to make an informed decision about the best keyframe which will reduce the summary distance. AAV-MMR results are slightly better than those of SAV-MMR. This is also expected, since SAV-MMR has the restriction that the audio and visual information have to originate from the same instant in the video, while AAV-MMR does not have this restriction. This restriction might be important in some cases, for example, if the summary contains the face image of a person, it might be preferable that the selected audio segment corresponds to this person speaking. However, in other cases, such as the description of a panorama, the synchronization between audio and video might not be of great importance. The results above show that the synchronization restriction increases the summary distance by an average of 25%. The actual choice of the most adequate method remains dependent on the intended application.

5.1.4 Conclusion of AV-MMR

In this section we have extended Video-MMR summarization algorithm into a new algorithm, AV-MMR, which combines both audio and visual information. The algorithm incrementally builds a summary by selecting segments which are similar to the whole content, but dissimilar to previously selected segments. We have proposed several variants of



(a) SRC of AM-AV-MMR



(b) SRC of GM-AV-MMR

Figure 5.2: SRC of AV-MMR

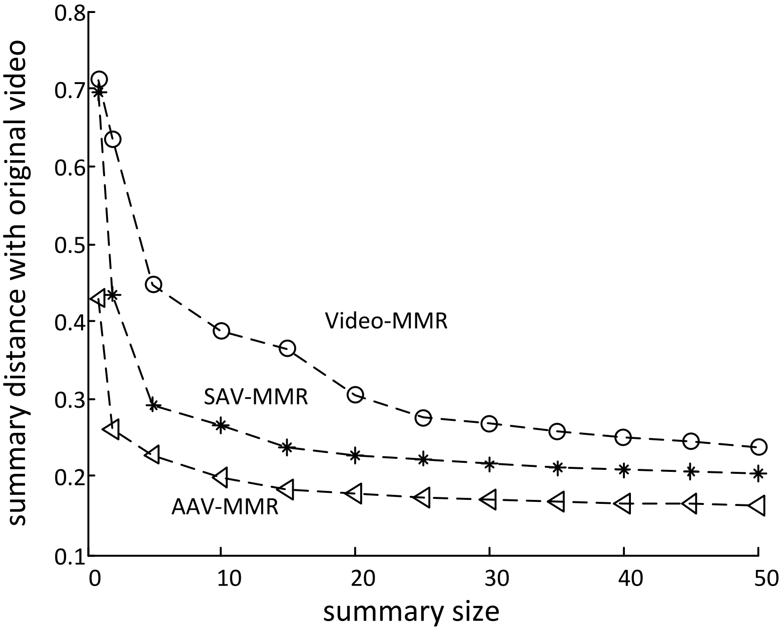


Figure 5.3: Comparison of Video-MMR and AV-MMR

the algorithm, for various definitions of the similarity measures, or depending on a synchronization constraint. Through experimentation, we have been able to select the best values for the parameters involved in the algorithm. The AM-AV-MMR variant appears to be the best one, so AV-MMR in the following sections refers to AM-AV-MMR. We have also shown that the AV-MMR algorithm produces better summaries than the previous Video-MMR, because AV-MMR is more similar to the original videos for visual and audio information.

5.2 Balanced Audio Video MMR

We have proposed multi-video summarization algorithms, Video-MMR only using visual information and AV-MMR exploiting both audio and visual information. However, AV-MMR is a simple extension of Video-MMR, and does not consider the specific characteristics of audio and video track. Therefore, in this section we propose an evolved algorithm, Balanced AV-MMR or BAV-MMR (Balanced Audio Video Maximal Marginal Relevance) [Li and Merialdo, 2012], to improve AV-MMR by:

- Considering the balance between audio information and visual information in a short time;
- Analyzing and using the influence of audio genres;
- Exploiting audio changes from one genre to another;
- Analyzing and utilizing the information brought by the face;
- Using the temporal distance of video frames in a set;
- Finally designing a novel mechanism to combine these features.

Table 5.1: The percentages of audio frames of each audio genre

Percentages of audio frames	Silence	Music	Speech
Document	0.0294	0.2732	0.6974
News	0.0298	0.2821	0.6881
Music	0.0259	0.2150	0.7591
Advertisement	0.0356	0.5726	0.3918
Cartoon	0.0205	0.4300	0.5495
Movie	0.0153	0.3933	0.5915
Sports	0.0508	0.4492	0.5000

The rest of this section is organized as follows: First we discuss the property of audio track and the importance of human face. And then the theory of Balanced AV-MMR is proposed to use these features and balance information. After that we present the experimental results of Balanced AV-MMR.

5.2.1 Analysis of audio track

Before exploiting audio information in Balanced AV-MMR it is necessary to analyze the characteristics of audio track. Here we analyze the audio through the genres. The audio can be classified into several genres: speech, music, speech&music, noise, silence and so on. The property of each genre is obviously different.

We consider one second as an atom, which we call ”audio frame”. Besides audio frame, contiguous audio frames with the same genre (silence, music or speech) are considered as an ”audio segment”. In the community of audio processing and speech recognition, Hidden Markov Model (HMM) is agreed to be a promising method. We use a successful toolkit, The Hidden Markov Model Toolkit (HTK) [University of Cambridge], to construct a recognition system of audio genres for audio frames. In this section we restrict audio genres to silence, music and speech because of the limitation of training data that we could annotate. Speech&music including singing is regarded as speech here. The test data is the audio tracks of 549 videos in 89 sets, which are downloaded from Wikio.fr same as Chapter 3, from 7 categories: Document, News, Music, Advertisement, Cartoon, Movie, and Sports. With these various audio files, we can guarantee the diversity of our audio files.

To analyze the audio frames and segments of each video category, we compute the percentage of silence, music and speech frames or segments in all the frames or segments of each video category. The percentages of audio frames of three genres are shown in Table 5.1, and Table 5.2 represents the percentages of audio segments. Since each category of video is analyzed separately, the sum of each row in Table 5.1 and Table 5.2 is 1.

In Table 5.1 and Table 5.2:

- *Sports* category obviously has the largest percentage of silence frames and segments compared with the other video categories. And the ratio of music segments to speech segments in *Sports* is high compared with the other categories.

- Compared with *Sports*, *Advertisement* category has a high percentage of silence segments but low percentage of silence frames, which means that silence segments are usually very short segments in *Advertisement*. And *Advertisement* contains short music segments and long speech segments.

Table 5.2: The percentages of audio segments of each audio genre

Percentages of audio segments	Silence	Music	Speech
Document	0.0554	0.4905	0.4541
News	0.0317	0.4869	0.4814
Music	0.0557	0.4629	0.4814
Advertisement	0.0979	0.4756	0.4265
Cartoon	0.0561	0.4619	0.4819
Movie	0.0530	0.4899	0.4571
Sports	0.1074	0.4862	0.4065

- *Advertisement* is definitely different from *Sports* according to above analysis.
- Refer to *Music* and *News*, they have similar percentages of three kinds of segments, but the percentages of the frames are different. The ratio of speech to music in *Music* category is larger compared to this ratio in *News*.

We only use several example to reveal the properties of different genres. Nevertheless, through the above analysis we can see that different categories of videos own obvious and different audio characteristics. The genres of audio frame and segments are indispensable features of the video.

Audio frames with the same genres seem to be more similar at the semantic level because of their same genre. Furthermore, the boundaries between audio segments, defined as "audio transition" in this thesis, are important because of the possible significant changes of visual information and audio information. For example in *News* the transition from music to speech genre is probably the beginning of the speech of an anchorman or journalist. Consequently, we will exploit audio genres and audio transitions in Balanced AV-MMR to improve the previous AV-MMR.

5.2.2 Analysis of human face

Human face is particularly important in video track, because most of current videos are human oriented. Moreover, the video track is relevant to the audio track and the appearance of the face in the video cannot be isolated with the audio track, so we carry out the analysis of the face in different audio genres.

We exploit the toolkit provided by Mikael Nilsson in [Nilsson et al., 2007] to detect the face in 89 video sets mentioned in Section 5.2.1. The percentage of frames with faces of each audio genre in all the frames is shown in Table 5.3 for each video category. Because there are possibly several faces in a frame, we also present the percentages of the number of faces of each audio genre in the total amount of faces in Table 5.4 for each video category. The sum of each row is equal to 1. As well, we make the analysis of large faces. The large face here is defined as the face with both the width and height larger than 90 pixels (video size is 320 by 240 pixels). The percentages of large faces in faces of each audio genre are shown in Table 5.5.

Comparing Table 5.3 to Table 5.4, the difference is small so the influence of the frames comprising multiple faces is small. And

- In video categories of *Document*, *News* and *Music*, most faces appear in speech audio frames. This is consistent with the characteristics of these videos, where the singer and reporter speak a lot.

Table 5.3: The percentage of frames with faces in the frames of each audio genre

Number of frames	Silence	Music	Speech
Document	0.0090	0.2003	0.7907
News	0.0027	0.2333	0.7640
Music	0.0039	0.1141	0.8820
Advertisement	0.0220	0.5291	0.4489
Cartoon	0.0053	0.3686	0.6262
Movie	0.0119	0.4685	0.5196
Sports	0.0313	0.3697	0.5990

Table 5.4: The percentages of faces of each audio genre in all the face

Number of faces	Silence	Music	Speech
Document	0.0094	0.2078	0.7828
News	0.0028	0.2330	0.7642
Music	0.0038	0.1104	0.8858
Advertisement	0.0240	0.5257	0.4502
Cartoon	0.0051	0.3784	0.6165
Movie	0.0090	0.3969	0.5941
Sports	0.0298	0.3761	0.5940

Table 5.5: The percentages of large faces in faces of each audio genre

Number of faces	Silence	Music	Speech
Document	0	0.2000	0.3151
News	0	0.1711	0.2921
Music	0.1000	0.0876	0.1874
Advertisement	0.2491	0.2352	0.1960
Cartoon	0.1667	0.1584	0.2269
Movie	0	0.0701	0.1543
Sports	0.2051	0.1362	0.1570

- In *Advertisement*, speech audio frames and music audio frames almost averagely share the number of faces because faces in *Advertisement* are uniformly distributed.
- In *Sports* up to around 3% faces are in silence audio frames as there are many human actions in silence while the other videos do not have similar characteristic.

In Table 5.5:

- 68% faces in *Advertisement* are large faces, indicating the characteristic of extremely human orientation. It is the same for *Cartoon* and *Sports* because of the existence of a large number of large faces. So a big spatial part of video frames is the face in *Cartoon*, *Sports* and *Advertisement.*
- In *News*, *Document* and *Movie*, there is not any large face in silence audio frames, caused by the silent prologue and epilogue containing few large faces.

The viewers of video summary - human beings favor the summary covering the significant frames with the face in the video. Moreover, we have only analyzed several characteristics of the face in some categories of videos, but it is obvious that the appearance of the face in a video is consistent with the category and property of this video. So the face is an important feature to improve our Balanced AV-MMR, and has strong relation with the audio track according to our analysis. Furthermore, a frame containing the face should be more similar to another frame with face than the frame without face in the semantic level.

5.2.3 Balanced AV-MMR

Assume that in a short time the audio attracted more attention from the user, the user would pay less attention to video content and vice versa, because the attention of a person in a short time is limited. In an audio segment, the duration is usually short. Therefore, there is a balance between audio information and visual information in an audio segment. Consequently we give our novel algorithm the name "balance". Balanced AV-MMR exploits the information from audio genre, the face and the time to improve the balance information and similarities of frames in semantic level.

According to the analysis in Section 5.2.1 and 5.2.2, audio genre and the face are important features in the video, which can influence the balance between audio and video in an audio segment. When audio transition happens, there is a significant change in the audio. At that time the user would pay more attention to the audio and the audio becomes more important than usual in the balance. Similarly, when the face appears in the video track of an audio segment, the video content becomes more important in the balance.

Moreover, the face and audio genre can influence the similarities between frames at the semantic level. For video track the similarity of two frames both containing the face is larger than the similarity between one frame with the face and another frame without the face. For audio track two frames from the same audio genre, for example the speech, are more similar.

In a video two close frames in time seem to be redundant. Two frames in a video seem less different than two frames from two individual and non-duplicated videos, even if they have the same similarities according to low-level features. Therefore it is necessary to consider the influence of temporal information on our summarization.

In this section, we will introduce several factors of audio, face and time to AV-MMR and propose the variants of Balanced AV-MMR.

5.2.3.1 Fundamental Balanced AV-MMR

From the formula of AV-MMR and the analysis of the balance between audio and video information in a segment, we introduce the balance factor between visual and audio information and generalize the fundamental formula of Balanced AV-MMR as:

$$\begin{aligned} f_{k+1} = \arg\max_{f \in V \backslash S_k} \{ & \rho(f)[\lambda Sim_{I1}(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g)] \\ & +(1-\rho(f))[\mu Sim_{A1}(f, A\backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)]\} \end{aligned} \tag{5.8}$$

In Section 5.1, it indicates $\lambda = 0.7$ and $\mu = 0.5$. Through bringing ρ into Eq. 5.8, Balanced AV-MMR considers the balance between audio and video. When ρ increases, the visual information takes a more important role in Balanced AV-MMR, and vice versa. Eq. 5.8 is our fundamental formula for the following variants. When ρ is equal to 0.5, Eq. 5.8 degenerates into AV-MMR.

5.2.3.2 Balanced AV-MMR V1: using audio genre

Through the audio analysis in Section 5.2.1, we have known that audio genre is an important feature and can reflect the characteristics of the videos. It is obvious that the audio frames with the same genre are more similar than the audio frames with different genres, even if they have the same similarity according to the audio features like Mel-frequency cepstral coefficients (MFCC). MFCC is used to compute the similarity of the short-term power property of two audio frames as AV-MMR in Section 5.1, but their similarity of semantic level cannot be reflected. Consequently, we can introduce an augment factor for audio genres to adjust the similarity of MFCC vectors. Here we use τ to denote this factor. Eq. 5.8 and its $Sim_{A1}(f, A\backslash S_k)$ and $Sim_{A2}(f,g)$ becomes:

$$\begin{aligned} f_{k+1} = \arg\max_{f \in V \backslash S_k} \{ & \rho(f)[\lambda Sim_{I1}(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g)] \\ & +(1-\rho(f))[\mu Sim'_{A1}(f, A\backslash S_k) - (1-\mu) \max_{g \in S_k} Sim'_{A2}(f,g)]\} \end{aligned} \tag{5.9}$$

and

$$\begin{aligned} & sim'_{A1}(f_i, A\backslash S_k) = \tfrac{1}{|A\backslash(S_k \bigcup f_i)|} \textstyle\sum_{f_j \in A\backslash(S_k \bigcup f_i)} \tau(f_i, f_j) sim(f_i, f_j); \\ & sim'_{A2}(f,g) = \tau(f,g) sim(f,g) \end{aligned} \tag{5.10}$$

where $sim(f_i, f_j)$ and $sim(f,g)$ are original similarities by MFCC, same with the definitions in Eq. 5.8. And $\tau(f_i, f_g) = 1 + \theta_\tau \cdot (\theta_P - |P(f_i) - P(f_g)|)$. θ_τ is a weight to adjust the influence of the audio genre. $\theta_P = 0.2$. $P(f_i)$ and $P(f_g) = 0$, 0.1, or 0.2 when the audio frame f_i is silence, music or speech genres.

Audio transitions indicate significant audio changes. In $Music$ category, the transition from silence or music audio to speech audio indicates the possible appearance of the singer, beginning singing at that time. In $News$ category, the transition from silence audio to speech audio usually indicates the start of the news by a journalist or an anchorperson.

Around audio transition the user would pay more attention to the audio and less attention to the video track, according to our balance principle. Consequently we modify the balance ratio ρ to ρ' by considering the transition factor φ_{tr}:

$$\rho'(f) = \frac{\rho(f)}{(\rho(f) + (1-\rho(f)) \cdot (1+\varphi_{tr}(f))} = \frac{\rho(f)}{(1+\varphi_{tr}(f) - \rho(f) \cdot \varphi_{tr}(f)} \tag{5.11}$$

Because of φ_{tr} and $\tau(f_i, f_j)$, the fundamental formula of Balanced AV-MMR, Eq. 5.9, transforms into the following formula, which is defined as the formula of Balanced AV-MMR V1:

$$\begin{aligned} f_{k+1} = \arg\max_{f\in V\backslash S_k} \{&\rho'(f)[\lambda Sim_{I1}(f, V\backslash S_k) - (1-\lambda)\max_{g\in S_k} Sim_{I2}(f,g)] \\ &+(1-\rho'(f))[\mu Sim'_{A1}(f, A\backslash S_k) - (1-\mu)\max_{g\in S_k} Sim'_{A2}(f,g)]\} \end{aligned} \tag{5.12}$$

5.2.3.3 Balanced AV-MMR V2: using face detection

According to the analysis in Section 5.2.2, the face is extremely important in visual information, so the video frame becomes more important when the face appears in a video frame. Since our balance principle favors one hand and dislikes the other hand in audio and visual information, the balance factor ρ' should increase in this case. After introducing the face factor β_{face} to $\rho'(f)$ in Section 5.2.3.2, it becomes:

$$\begin{aligned} \rho''(f) &= \frac{\rho(f)\cdot(1+\beta_{face}(f))}{\rho(f)\cdot(1+\beta_{face}(f)) + (1-\rho(f))\cdot(1+\varphi_{tr}(f))} \\ &= \frac{\rho(f)\cdot(1+\beta_{face}(f))}{1+\varphi_{tr}(f)+\rho(f)\cdot(\beta_{face}(f)-\varphi_{tr}(f))} \end{aligned} \tag{5.13}$$

where $\beta_{face}(f) = 1 + facenumber(f) \times \theta_{face}$. θ_{face} is a weight for adjusting the influence of the face.

Besides the balance factor $\rho''(f)$, the appearance of face also influences the similarity of two video frames. At the semantic level, a frame comprising face is more similar to another frame with face than to the frame without face. Also, two frames with faces often reveal the relevant content of the video, such as several journalists in *News* and actors in *Movie*. Therefore the similarities Sim_{I1} and Sim_{I2} in Eq. 5.8 evolve into:

$$\begin{aligned} &Sim'_{I1}(f_i, V\backslash S_k) = \tfrac{1}{|V\backslash(S_k \bigcup f_i)|}\textstyle\sum_{f_j\in V\backslash(S_k\bigcup f_i)} \beta'_{face}(f_i,f_j) sim(f_i,f_j) \\ &Sim'_{I2}(f,g) = \beta'_{face}(f,g)\cdot sim(f,g) \end{aligned} \tag{5.14}$$

where $\beta'_{face}(f_i, f_j) = 1 + (facenumber(f_i) + facenumber(f_j))/2 \times \theta_{face}$.

Based on above development, Eq. 5.12 of Balanced AV-MMR V1 can be reformulated as:

$$\begin{aligned} f_{k+1} = \arg\max_{f\in V\backslash S_k} \{&\rho''(f)[\lambda Sim'_{I1}(f, V\backslash S_k) - (1-\lambda)\max_{g\in S_k} Sim'_{I2}(f,g)] \\ &+(1-\rho''(f))[\mu Sim'_{A1}(f, A\backslash S_k) - (1-\mu)\max_{g\in S_k} Sim'_{A2}(f,g)]\} \end{aligned} \tag{5.15}$$

5.2.3.4 Balanced AV-MMR V3: adding temporal distance factor

At last, we prefer considering the influence of temporal distance of two frames f_i and f_j, from the same video or not, on the visual and audio similarities:

- Frames closer in time in a video commonly represent more relevant content, so two frames closer in a video are regarded more similar than two frames further in a video.
- For multiple videos, a frame is more similar to another frame in the same video than a frame from another non-duplicated video.

Then we can consider temporal information for selecting frames from multiple videos to the summary. This balance is called ”temporal balance”. The temporal factor is named as α_{time} and

$$\alpha_{time}(f_i, f_j) = \begin{cases} 1, & \text{if } f_i \text{ and } f_j \text{ are from two videos;} \\ 1 + \theta_{time} \cdot (1 - \frac{|t(f_i) - t(f_j)|}{10 * D_M}), & \text{if } f_i \text{ and } f_j \text{ are from the same video.} \end{cases} \tag{5.16}$$

where $t(f_i)$ and $t(f_j)$ are the frame times of f_i and f_j in video M. D_M is the total duration of video M. θ_{time} is a weight to adjust the influence of the temporal distance. Then the similarities of the frames in Balanced AV-MMR become:

$$\begin{aligned} &Sim''_{I1}(f_i, V\backslash S_k) = \tfrac{1}{|V\backslash(S_k \bigcup f_i)|} \textstyle\sum_{f_j \in V\backslash(S_k \bigcup f_i)} \beta'_{face}(f_i, f_j)\alpha_{time}(f_i, f_j) sim(f_i, f_j) \\ &Sim''_{I2}(f, g) = \beta'_{face}(f, g)\alpha_{time}(f_i, f_j) sim(f, g) \\ &sim''_{A1}(f_i, A\backslash S_k) = \tfrac{1}{|A\backslash(S_k \bigcup f_i)|} \textstyle\sum_{f_j \in A\backslash(S_k \bigcup f_i)} \tau(f_i, f_j)\alpha_{time}(f_i, f_j) sim(f_i, f_j) \\ &sim''_{A2}(f, g) = \tau(f, g)\alpha_{time}(f_i, f_j) sim(f, g) \end{aligned} \tag{5.17}$$

Consequently, the formula of Balanced AV-MMR V3 is similar to Eq. 5.15 of Balanced AV-MMR V2 and generalized as

$$\begin{aligned} f_{k+1} = \mathop{\arg\max}_{f \in V\backslash S_k} \{ &\rho''(f)[\lambda Sim''_{I1}(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim''_{I2}(f, g)] \\ &+ (1 - \rho''(f))[\mu Sim''_{A1}(f, A\backslash S_k) - (1-\mu) \max_{g \in S_k} Sim''_{A2}(f, g)] \} \end{aligned} \tag{5.18}$$

5.2.3.5 The procedure of Balanced AV-MMR

In the above sections, we have explained the formulas of fundamental Balanced AV-MMR, Balanced AV-MMR V1, Balanced AV-MMR V2 and Balanced AV-MMR V3. We need to generalize the procedure of Balanced AV-MMR like AV-MMR:

1. Detect the audio genres of the frames by HTK audio system described in Section 5.2.1, and the face by the toolkit in Section 5.2.2;

2. Compute importance ratio ρ, ρ', or ρ'' for each audio segment;

3. The initial video summary S_1 is initialized with one frame, defined as:

   $$S_1 = \mathop{\arg\max}_{f_i, f_i \neq f_j} [\prod_{j=1}^{n} Sim_I(f_i, f_j) \cdot \prod_{j=1}^{n} Sim_A(f_i, f_j)]^{\frac{1}{n}} \tag{5.19}$$

   where f_i and f_j are frames from the set V of all frames from all videos, and n is the total number of frames except f_i. Sim_I computes the similarity of image information between f_i and f_j; while Sim_A is the similarity of audio information between f_i and f_j;

4. Select the frame f_k by the formula of a variant of Balanced AV-MMR;

5. Set $S_k = S_{k-1} \bigcup \{f_k\}$;

6. Iterate to step 4 until S has reached the predefined size.

5.2.4 Experimental results of Balanced AV-MMR

Our experimental videos for Balanced AV-MMR are 36 video sets from 7 categories mentioned in Section 5.2.1, comprising 194 videos. Each video set contains 3-15 videos, each of which has the duration of 10 seconds to more than 10 minutes. The diversity of our experimental videos ensures the generic property of the summary produced by BAV-MMR.

The visual content of a keyframe is represented by the Bag-Of-Word (BOW) feature. BOW feature vector of a keyframe is the histogram of the number of visual words that appear in the keyframe. The similarity between two keyframes $sim(f_i, f_j)$ is computed like Eq. 3.2 as,

$$sim_I(f_i, f_j) = cos(H_{f_i}, H_{f_j}) = \frac{H_{f_i} \cdot H_{f_j}}{\|H_{f_i}\| \cdot \|H_{f_j}\|} \tag{5.20}$$

where H_{f_i} and H_{f_j} are the visual word histograms of keyframes f_i and f_j. Audio feature uses MFCC obtained by SPro Toolkit [IRISA]. The similarity of two averaged MFCC vectors is computed and normalized as

$$sim_A(a_i, a_j) = 1 - \frac{|a_i - a_j|}{\max_{a_m, a_n \in S_{MFCC}}(|a_m - a_n|)} \tag{5.21}$$

where a_i, a_j, a_m and a_n are averaged MFCC vectors.

To verify the effect of BAV-MMR, we use Audio Video Distance (AVD) and Video Distance (VD) of the summary with the original videos. AVD is defined as

$$d_{AVD}(S, V) = \frac{1}{n}\sum_{j=1}^{n} \min_{f_j \in V, g \in S}[1 - (sim_I(f_j, g) + sim_A(f_j, g))/2] \tag{5.22}$$

where n is the number of frames in V. g and f_j are frames respectively from video summary S and V. And similarly VD is defined as

$$d_{VD}(S, V) = \frac{1}{n}\sum_{j=1}^{n} \min_{f_j \in V, g \in S}(1 - sim_I(f_j, g)) \tag{5.23}$$

In the fundamental formula of BAV-MMR, Eq. 5.8, we need to decide the value of parameter ρ. In this section we consider ρ as a constant value for different frames. To remain consistent with Video-MMR and AV-MMR, we use the same method, Summary Reference Comparison (SRC), comparing the summary qualities from different weights, to decide ρ. ρ varies from 0.0 to 1.0, with each step of 0.1. The results of SRC are shown in Fig. 5.4. SRC here uses $d_{AVD}(S, V)$.

From Fig. 5.4 when $\rho = 0$, $d_{AVD}(S, V)$ is large when the summary size is small and vice versa. Since we want a commonly small $d_{AVD}(S, V)$ for different summary sizes, at last we select $\rho = 0.5$.

By trial and error, the various parameters in the variants of BAV-MMR are set to the following values:

- In Eq. 5.11 $\varphi_{tr}(f) = 0.1$ when the audio transits from silence to music at f and vice versa, or from speech to music and vice versa; $\varphi_{tr}(f) = 0.2$ when the audio transits from silence to speech and vice versa; and $\varphi_{tr}(f) = 0$ if there is not any audio transition in frame f.

- The weights θ_τ, θ_{time} and θ_{face} are chosen as 0.3, 0.3 and 0.2.

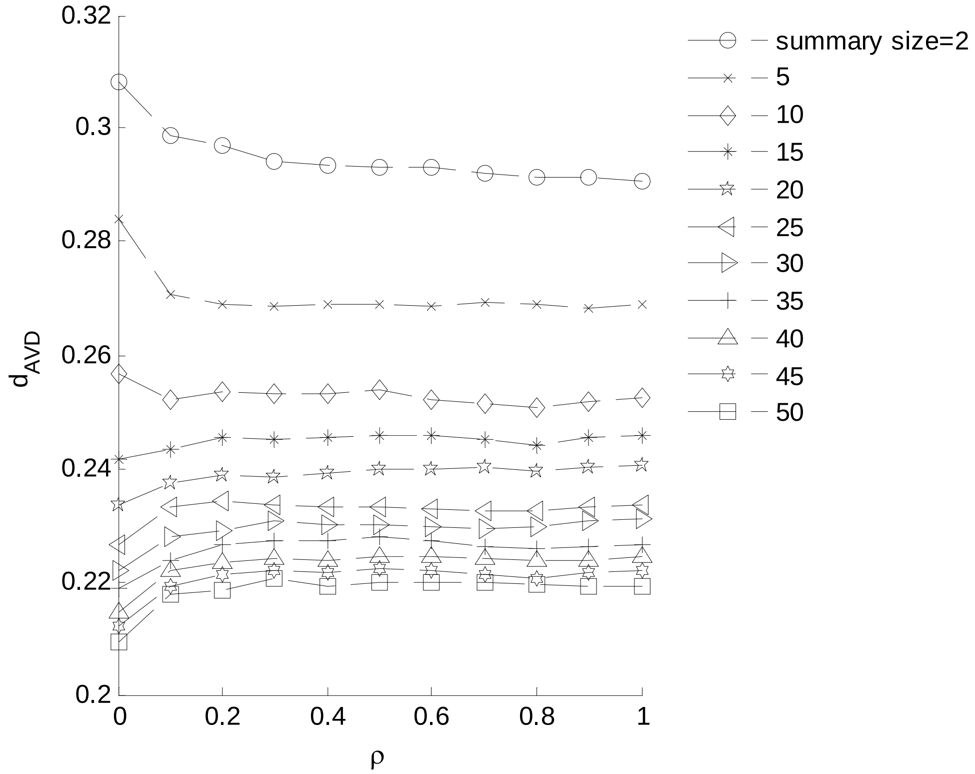


Figure 5.4: SRC of ρ_T

The means of AVDs and VDs of 36 experimental video sets from Video-MMR, AV-MMR and the variants of BAV-MMR are shown in Fig. 5.5 and Fig. 5.6. We have not shown the curve of the fundamental BAV-MMR with $\rho = 0.5$ which is the same with AV-MMR in Fig. 5.4. It is clear that the variants of BAV-MMR are better than Video-MMR and AV-MMR because of the smaller distances with the original videos. Among the variants of BAV-MMR, BAV-MMR V1 is better than AV-MMR, and BAV-MMR V2 is better than BAV-MMR V1. While BAV-MMR V3 outperforms BAV-MMR V2 a lot because BAV-MMR V3 improves the algorithm in both audio and video track, but BAV-MMR V1 and BAV-MMR V2 separately improve audio track and video track in the summarization.

When the summary size increases, the improvements of BAV-MMR V1 and BAV-MMR V2 is not as good as the smaller summary size, which is caused by more various audio genres and more face types in video summaries. However, the temporal information is not influenced a lot by the selected frames in the summaries, so BAV-MMR V3 keeps its curve trend when the summary size increases.

The limitation of BAV-MMR is the manual decisions of the weights φ_{tr}, θ_{τ}, θ_{time} and θ_{face}. So it is necessary to automatically and optimally tune these weights for a generic summarization algorithm. A particular set of optimized weights for each category of video is favorable. Furthermore, BAV-MMR may benefit from a variable ρ according to the property of frame or segment.

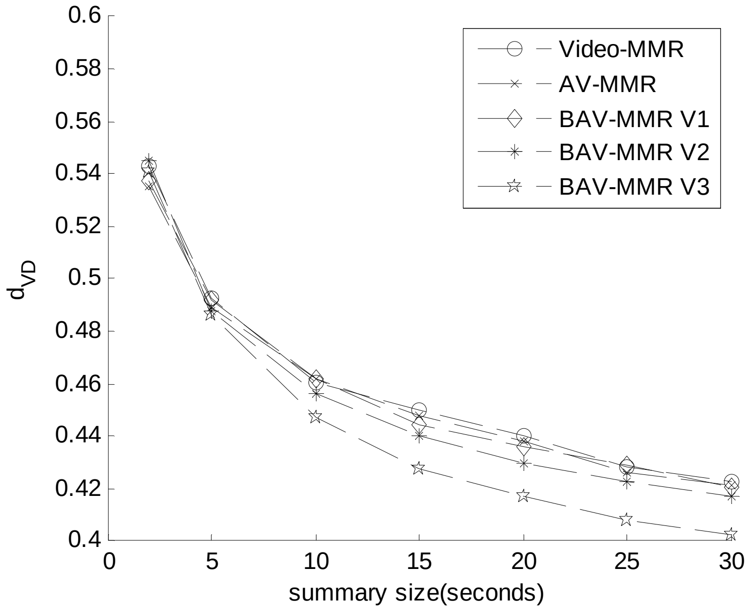


Figure 5.5: SRC of different measures for BAV-MMR by VD

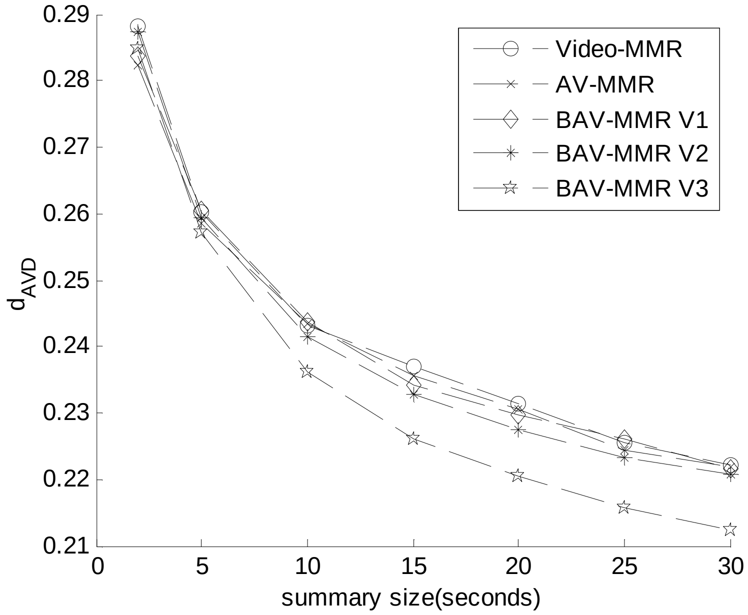


Figure 5.6: SRC of different measures for BAV-MMR by AVD

5.2.5 Conclusion of Balanced AV-MMR

We have proposed a novel multi-video summarization algorithm, Balanced AV-MMR by considering the balance between audio and visual information in a segment, and temporal balance of inter- and intra- video. Besides, we use audio genre and the face to adjust the similarities of the frames. Balanced AV-MMR is a new improvement of the series of MMR algorithms in video summarization. And several variants of BAV-MMR have been proposed and proved better than previous algorithms. However, there are still some limits for Balanced AV-MMR. The weights in Balanced AV-MMR are manually decided, but it is likely that they could benefit from the adaptation to the video genre. So it is necessary to automatically optimize the weights to the category of the video, summary size, and so on.

5.3 Optimized Balanced Audio Video MMR

5.3.1 The principle of OB-MMR

In the previous sections, we have proposed a series of generic algorithms, Video-MMR for multi-video summarization by using only visual information, AV-MMR by exploiting both audio and visual information, and Balanced AV-MMR which considers the balance factor between audio and visual information. In this section, we improve over Balanced AV-MMR by optimizing the parameters in the algorithm, and propose OB-MMR (Optimized Balanced Audio Video MMR) [Li and Merialdo, 2011].

In Section 5.2 Balanced AV-MMR exploits many parameters which are manually set according to experience. These parameters are: the balance parameter between audio and visual information ρ''_T, the parameter of temporal distance α_{time}, the face parameter β'_{face}, the audio genre parameter τ, and the parameter for audio transition φ_{tr}.

However, it is hard to manually decide the best values of these 5 parameters. Also for different genres of videos, the optimal values of those parameters may vary, because the relation between video track and audio track is different. Therefore, we wish to propose an automatic mechanism to optimize the set of weights for Balanced AV-MMR. The first thing is to reformulate Eq. 5.18 into the following formula:

$$\begin{aligned} f_{k+1} = \arg\max_{f\in V\backslash S_k}\{&\rho''(f)[\lambda Sim''_{I1}(f,V\backslash S_k) - (1-\lambda)\max_{g\in S_k} Sim''_{I2}(f,g)] \\ &+(1-\rho''(f))[\mu Sim''_{A1}(f,A\backslash S_k) - (1-\mu)\max_{g\in S_k} Sim''_{A2}(f,g)]\} \end{aligned} \tag{5.24}$$

where

$$\rho''_T(f) = \frac{B(f)\cdot W_b + F(f)\cdot W_f}{(B(f)\cdot W_b+F(f)\cdot W_f)+((1-B(f))\cdot W_b+R(f)\cdot W_r)};$$

$$Sim''_{I1}(f_i,V\backslash S_k) = \frac{1}{|V\backslash(S_k\bigcup f_i)|}\sum\nolimits_{f_j\in V\backslash(S_k\bigcup f_i)}[1+(F(f_i)+F(f_j))\cdot W_f] \\ (1+T(f_i,f_j)\cdot W_t)sim(f_i,f_j);$$

$$Sim''_{I2}(f,g) = [1+(F(f)+F(g))\cdot W_f](1+T(f,g)\cdot W_t)sim(f,g); \tag{5.25}$$

$$Sim''_{A1}(f_i,A\backslash S_k) = \frac{1}{|A\backslash(S_k\bigcup f_i)|}\sum\nolimits_{f_j\in A\backslash(S_k\bigcup f_i)}(1+S(f_i,f_j)\cdot W_s) \\ (1+T(f_i,f_j)\cdot W_t)sim(f_i,f_j);$$

$$Sim''_{A2}(f,g) = (1+S(f,g)\cdot W_s)(1+T(f,g)\cdot W_t)sim(f,g)$$

Inside Eq. 5.24, the functions B, F, T, R, and S are the computed features for the balance between audio and visual information, the face importance, the temporal distance, the audio transition, and the audio genre; while, W_b, W_f, W_t, W_r, and W_s are the weights for those features B, F, T, R, and S. Compared to Eq. 5.18, Eq. 5.24 is easier to be optimized because we just have to automatically adjust the values of the weights W_b, W_f, W_t, W_r, and W_s to achieve the best result. The result of the optimization of Eq. 5.24 is called OB-MMR.

Before adjusting the weights in Eq. 5.18, we need to define the fitness function for these weights. In video summarization, we usually regard the summaries from human being as the ground truth, because video summarization is a problem which is absolutely human oriented. Assume that we already have some groups of human summaries, we could use the similarities between the summaries from OB-MMR and the summaries from human as the fitness function to adjust the weights W_b, W_f, W_t, W_r, and W_s as the fitness function, because we want the summary from OB-MMR more similar to the summary from human.

Then it is necessary to select an automatic algorithm to automatically tune the weights W_b, W_f, W_t, W_r and W_s. One successful algorithm is Particle Swarm Optimization (PSO) proposed by [Poli et al., 2007] [Kennedy and Eberhart, 1995]. PSO has been used across a wide range of applications, which has proved the effect of PSO. ”The swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function. A particle by itself has almost no power to solve any problem ; progress occurs only when the particles interact.” [Poli et al., 2007]. In PSO, every particle decides its movement by considering its current location and the previous best location of the particles. The individual contains three D-dimensional vectors: the current position $\overrightarrow{x_i}$, the previous best position $\overrightarrow{p_i}$, and the velocity $\overrightarrow{v_i}$. The best function result is denoted by $pbest_i$, and $\overrightarrow{p_g}$ is the best neighbor of $\overrightarrow{p_i}$. PSO procedure is described in [Poli et al., 2007]. In OB-MMR, 5 weights W_b, W_f, W_t, W_r, and W_s can be considered as 5 elements of a vector $\overrightarrow{x_i}$ in the searching space of PSO for the fitness function. The process for implementing PSO described in [Poli et al., 2007] is as in Table 5.6.

Another simple algorithm to find the optimized weights for OB-MMR is the Gridding and Relaxation (GR). Here the gridding means averagely gridding the possible weights in a suitable range, and trying every combination of the weights to optimize the fitness function, the similarity between OB-MMR summary and human summary. Since the interval between two gridding values is initially large so that the computation is fast enough, when the best values are found on the grid, a similar process is repeated recursively with a finer grid around the optimal point, which is called the relaxation step. The fitness function for gridding and relaxation is the same with above, the similarity between OB-MMR summary and human summary. The relaxation occurs multiple times, until the grid interval reaches the desired precision.

The optimized weights from PSO and GR will be shown in the next section. After we obtain the optimized weights, we could implement OB-MMR process in a way similar to Balanced AV-MMR, shown in Table 5.7. To better explain the structure, we illustrate the framework in Fig. 5.7.

5.3.2 Experimental results of OB-MMR

We have two video sets, ”DATI” and ”YSL” as Video-MMR in Section 3.2.3, which are both news videos obtained from the aggregation website, ”WIKIO”. There are 16 videos in ”DATI”, and 14 videos in ”YSL”. Both video sets own videos with the duration from around

Table 5.6: Algorithm: Particle Swarm Optimization

1. Initialize a population array of particles with random positions and velocities on D dimensions in the search space.

2. **Loop**

3. For each particle, evaluate the desired optimization fitness function in D variables.

4. Compare particle's fitness evaluation with its $pbest_i$. If current value is better than $pbest_i$, then set $pbest_i$ equal to the current value, and $\overrightarrow{p}_i$ equal to the current location $\overrightarrow{x}_i$ in D-dimensional space.

5. Identify the particle in the neighborhood with the best success so far, and assign its index to the variable g.

6. Change the velocity and position of the particle according to the following equation:

   $$\begin{cases} \overrightarrow{v}_i \leftarrow w\overrightarrow{v}_i + \overrightarrow{U}(0,\phi_1) \bigotimes (\overrightarrow{p}_i - \overrightarrow{x}_i) + \overrightarrow{U}(0,\phi_2) \bigotimes (\overrightarrow{p}_g - \overrightarrow{x}_i) \\ \overrightarrow{x}_i \leftarrow \overrightarrow{x}_i + \overrightarrow{v}_i \end{cases} \tag{5.26}$$

   where $\overrightarrow{U}(0,\phi_2)$ represents a vector of random numbers uniformly distributed in $[0,\phi_i]$ which is randomly generated at each iteration and for each particle.

7. If a criterion is met (usually a sufficiently good fitness or a maximum number of iterations), exit loop.

8. **End loop**

Table 5.7: Algorithm: OB-MMR

1. Summarize video track by Video-MMR with Eq. 3.11.
2. Summarize the audio track by Audio-MMR:

$$S_{k+1} = S_k \bigcup \mathop{\arg\max}_{f \in A\backslash S_k} (1 + S(f,g) \cdot W_s) \cdot (\lambda(Sim_1(f, A\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_2(f,g)) \tag{5.27}$$

3. Detect the audio segments and their genres by HTK audio system.
4. The initial video summary S_1 is initialized with one frame, defined as:

$$S_1 = \mathop{\arg\max}_{f_i, f_i \neq f_j} [\prod_{j=1}^{n} Sim_I(f_i, f_j) \prod_{j=1}^{n} Sim_A(f_i, f_j)]^{\frac{1}{n}} \tag{5.28}$$

   where f_i and f_j are frames in video set V, and n is the total number of frames except f_i. Sim_I computes similarity of image information between f_i and f_j, while Sim_A is the similarity of audio information between f_i and f_j.

5. Find the optimized weights for Eq. 5.24 by fitting the OB-MMR summary to human summaries.
6. **Loop**
7. Select frame f_{k+1} by Eq. 5.24 with optimized weights.
8. Set $S_{k+1} = S_k \bigcup \{f_{k+1}\}$.
9. Iterate to Step 7 until S has reached the predefined size.
10. **End loop**

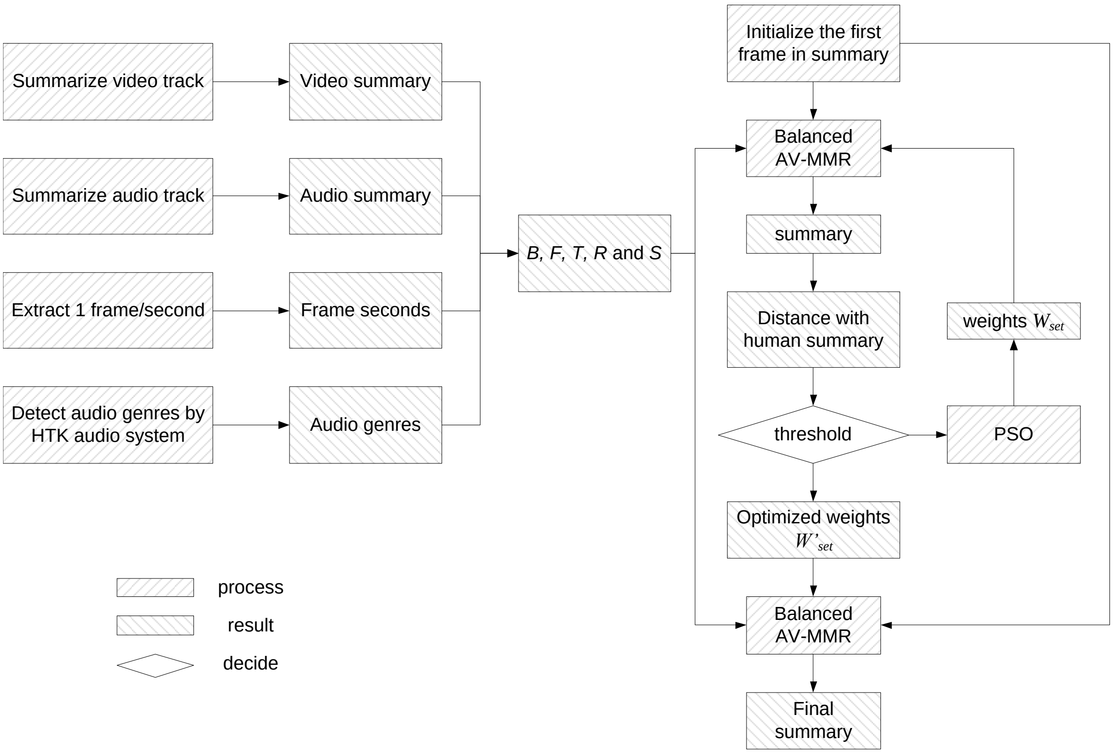


Figure 5.7: The framework of OB-MMR

Table 5.8: The number of audio frames with different genres

	silence	music	speech
DATI	24	524	2366
YSL	57	1173	1318

30 seconds to around 10 minutes, and video categories vary from *News*, *Advertisements* and *Music* video to *Movie*.

The visual content of a keyframe is represented by the Bag-Of-Word feature vectors [Li and Merialdo, 2010e] [Video Retrieval Group, City U. of Hong Kong], and audio feature are MFCC vectors obtained by SPro Toolkit [IRISA]. The similarities between audio features and visual features are identical to their definition in AV-MMR. And visual similarity is used as the fitness function in OB-MMR.

As Balanced AV-MMR, we also use the trained HTK toolkit [University of Cambridge] to process the audio track and get the audio genre of each audio frame. The statistical data of audio genres of audio frames in ”DATI” and ”YSL” is shown in Table 5.8. The toolkit provided in [Nilsson et al., 2007] is used to detect faces in the video frames. Furthermore, in Section 6.3 we get the human summaries for ”DATI” and ”YSL”, which are used in the fitness function of PSO and GR to get the optimized weights for OB-MMR. The human summaries used as the ground truth are assessed by 12 people with professional background of image processing. Each person selects 10 most important video frames for each video in our prepared frames. In PSO, we consider a population, individual sets of weights, of size 20. And according to [Poli et al., 2007] $w = 0.7298$, and $\phi_1 = \phi_2 = 1.49618$ in Eq. 5.24. In the gridding, the range of gridding for each weight is [0.0,2.0] and the

Table 5.9: Optimized weights and similarities

	PSO		GR	
	DATI	YSL	DATI	YSL
similarity	0.35410	0.33616	0.34991	0.31470
W_b	0.65946	0.69707	0.75000	0.75000
W_f	0.22887	0.01958	0.25000	0.25000
W_t	1.50302	0.29761	1.75000	0.25000
W_r	0.13954	0.09365	0.25000	0.25000
W_s	0.08224	0.21695	0.05000	0.63000

Table 5.10: Cross validation

	Weights from PSO		Weights from GR	
	DATI	YSL	DATI	YSL
similarity	0.35410	0.33616	0.34991	0.31470
DATI	0.35410	0.19519	0.34991	0.19350
YSL	0.27719	0.33616	0.26851	0.31470

initial interval of gridding is 0.5. During the relaxation phrases, the intervals of the two relaxation iterations are 0.04 and 0.006 respectively.

And in Eq. 5.24 of OB-MMR, we define the parameters T, F, B, R and S as follows:

$$T(f_i, f_j) = \begin{cases} 1, & \text{if } f_i \text{ and } f_j \text{ are from two videos;} \\ 1 + 0.1 \cdot (1 - \frac{|t(f_i)-t(f_j)|}{10*T_M}), & \text{if } f_i \text{ and } f_j \text{ are from the same video.} \end{cases} \tag{5.29}$$

where $t(f_i)$ and $t(f_j)$ are time orders of f_i and f_j in video M, which owns a time duration T_M;

$$F(f) = \textit{face number in frame } f; \tag{5.30}$$

$$B(f) = N_V(f)/(N_V(f) + N_A(f)), \tag{5.31}$$

where $N_V(f)$ is frame number of video summary, created by Video-MMR, in audio segment L where frame f is inside, and $N_A(f)$ is frame number of audio summary, by Audio-MMR, in L;

$$R(f_i, f_{i+1}) = |k(f_i) - k(f_{i+1})|, \text{where } k(f) = \begin{cases} 0.1, & \text{frame } f \text{ is silence audio frame,} \\ 0.3, & \text{frame } f \text{ is music audio frame,} \\ 0.4, & \text{frame } f \text{ is speech audio frame.} \end{cases} \tag{5.32}$$

$$S(f_i, f_j) = |m(f_i) - m(f_j)|, \text{where } m(f) = \begin{cases} 0.5, & \text{frame } f \text{ is silence audio frame,} \\ 0.8, & \text{frame } f \text{ is music audio frame,} \\ 0.9, & \text{frame } f \text{ is speech audio frame.} \end{cases} \tag{5.33}$$

The optimized weights of W_b, W_f, W_t, W_r, and W_s by PSO and GR and their corresponding similarities of fitness function are shown in Table 5.9.

To prove the effect of the optimized weights from ”DATI” and ”YSL”, the cross validation is used here, which means that the weights from ”DATI” are used to compute the similarity of ”YSL” and vice versa. The results of cross validation are shown in Table 5.10.

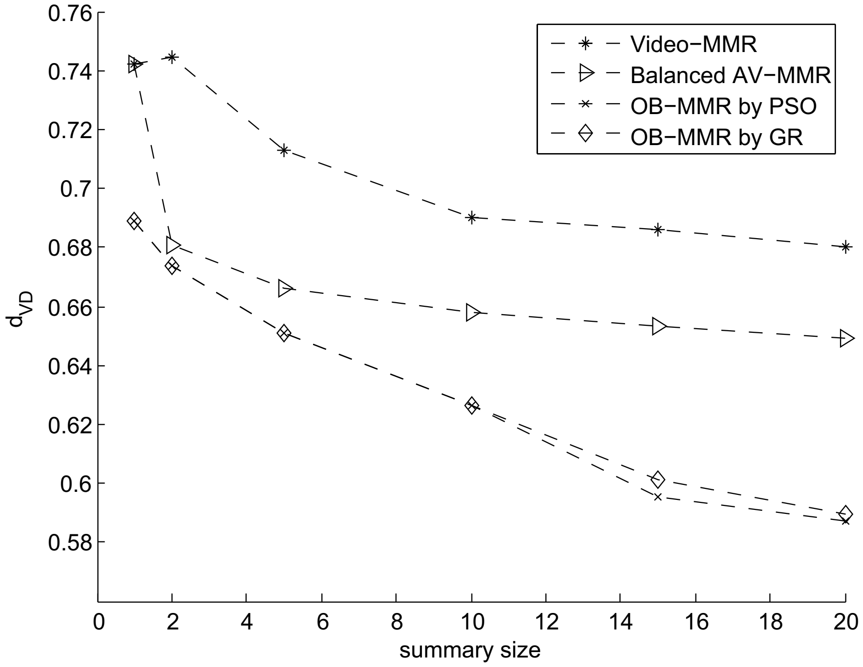


Figure 5.8: SRC of different measures for OB-MMR by VD

From Table 5.10, it is obvious that the weights from PSO and GR for ”DATI” are better than the weights from PSO and GR for ”YSL” for both video sets, while the similarities are very similar in Table 5.9. Therefore we exploit these two sets of optimized weights, from PSO and GR, of ”DATI” other than ”YSL” in the following experiments.

Then we use these 2 sets of weights to compute Video Distance (VD) and Audio Video Distance (AVD) between OB-MMR summaries and original videos like Balanced AV-MMR in Section 5.2.

The results are shown in Fig. 5.8 and Fig. 5.9. It is obvious that two OB-MMR curves are better than the previous algorithms, Video-MMR and Balanced AV-MMR. OB-MMR by PSO is a little better than OB-MMR by GR, which is caused by the better similarity with the human summaries shown in Table 5.9. And even when the optimized weights from video set ”DATI” are used in OB-MMR for ”YSL”, the results in Fig. 5.8 and Fig. 5.9 are better. So OB-MMR is a generic algorithm, even the optimized weights are from the other videos. In the future, it is possible to decide a fixed optimized set of weights for each genre of videos after the large-scale experiments.

We could also conclude that PSO is better than GR for OB-MMR according to the curves shown in Fig. 5.8 and Fig. 5.9. Furthermore, it is unnecessary for PSO to define the range and intervals to do the gridding and relaxation before the computation, so PSO is an unsupervised algorithm and better for OB-MMR. OB-MMR optimized by PSO can be used to summarize different categories of videos without a prior knowledge except the video category.

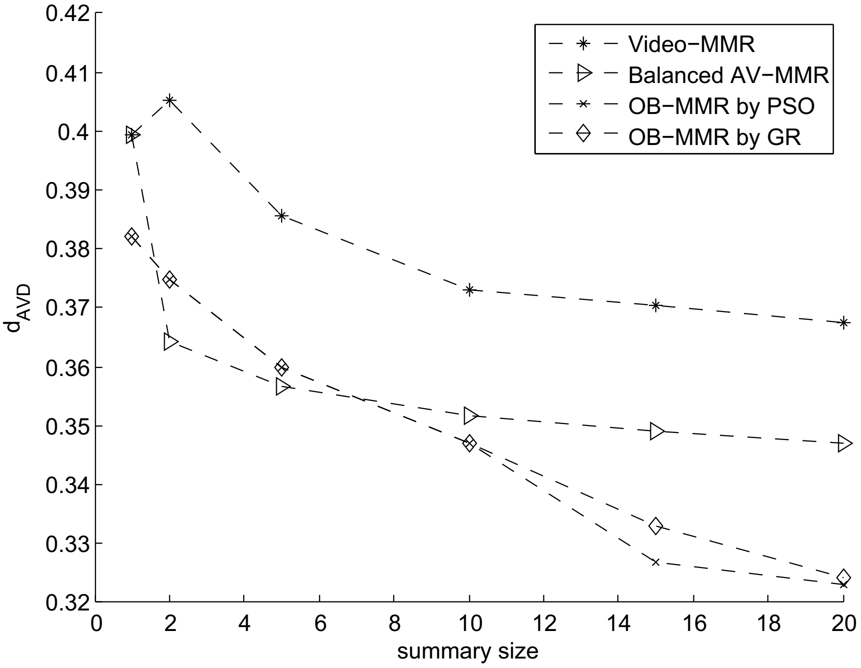


Figure 5.9: SRC of different measures for OB-MMR by AVD

5.3.3 Conclusion of OB-MMR

In this section, we have proposed a summarization algorithm, OB-MMR, which better resolves the problem of combining audio and visual information during the summarization than previous algorithms, and is able to summarize multi-video. OB-MMR improves its predecessor, Balanced AV-MMR, by automatically adjusting the optimized weights fitting to the known human summaries. But similar to Balanced AV-MMR, OB-MMR exploits several typical features in the video: temporal information, face, audio genre, and audio transition of the genre. In the same category of videos, even the optimized weights are from the other videos, OB-MMR could obtain a better summary than Video-MMR and Balanced AV-MMR. And between OB-MMR by PSO and OB-MMR by GR, PSO is the better one, because the summary from OB-MMR is more similar to the original video, and PSO does not need the prior knowledge of the video, like the range and interval of possible weights.

OB-MMR can use the same optimized weights for different categories of videos, but it is better for OB-MMR to decide one set of optimized weights for each category of video, such as news, movie, sports, and so on, by fitting the weights to the known human summaries, which needs large-scale human assessments. Consequently the next step of OB-MMR is to test and decide the optimized weights for different categories of videos by the large-scale experiments with massive video sets.

5.4 Text Video MMR

In previous sections, we have exploited the multimedia indices, video and audio, to improve the summarization. However, text information is also important in the video. Therefore,

Table 5.11: Some examples of 3-gram

Score	Begin	End	3-gram
0.06	51.53	52.18	on craint on
0.07	51.58	52.28	craint on s'
0.07	51.94	52.86	on s' exprimer
0.15	52.25	53.44	s' exprimer comédien
0.15	52.46	53.54	exprimer comédien ce
0.15	52.86	53.94	comédien ce matin
0.07	53.47	54.21	ce matin les

we propose Text Video MMR (TV-MMR) to exploit text and visual information to summarize the video. First we describe the method of obtaining the text utterance.

5.4.1 Linguistic information measure

In our approach, the information content of the audio track is evaluated based on the text transcription of the audio channel by an Automatic Speech Recognition (ASR) system from LIA (Laboratoire d'Informatique d'Avignon, France). The LIA ASR system is using context-dependent Hidden Markov Models for acoustic modeling and n-gram Language Models (LM). Training corpora comes from broadcast news records and large textual materials: acoustic models are estimated on 180 hours of French broadcast news. Language Models are trained on a collection of about 109M words, from French newspapers and large newswire collections. The ASR system is run on the audio track of the video sequences. The result is a sequence of words, with the beginning and ending times of their utterance. These timecodes allow synchronizing the audio and the video information in the summarization algorithm. They also allow providing candidate boundaries for audio-visual segments to be selected.

By analogy with text information retrieval techniques, the audio information content is measured according to the words that appear in the selected segment. We construct a word document vector d for the whole transcription of a video (or the transcriptions of a set of videos), as in the Vector Space model. We construct a similar vector for the text transcription t of a segment extracted from an audio-visual sequence. The audio information content of the segment is defined as the cosine between these two vectors:

$$sim(t,d) = cos(t,d) = \frac{t \cdot d}{\|t\|\|d\|} \tag{5.34}$$

The results are provided as lists of sliding windows of n words, (with n ranging from 1 to 10), together with windows covering complete sentences. For each window, the beginning and end times are provided, together with the similarity score. An example of such lists for 3-grams is shown in Table 5.11.

5.4.2 TV-MMR

TV-MMR selects video segments corresponding to n-grams by using both the textual and the visual content. By mimicking the formula of Video-MMR, the formula of TV-MMR is proposed as:

$$S_{k+1} = S_k \bigcup \underset{f \in V \backslash S_k}{\arg\max}\{\beta[\lambda Sim_{I1}(f, V \backslash S_k) - (1-\lambda)\max_{g \in S_k} Sim_{I2}(f,g)] + (1-\beta)[\mu Sim_{T1}(f, V \backslash S_k) - (1-\mu)\max_{g \in S_k} Sim_{T2}(f,g)]\} \tag{5.35}$$

where f and g are audio-visual segments corresponding to n-grams. The definitions of Sim_{I1} and Sim_{I2} are the same as in Eq. 3.11. Sim_{I1} and Sim_{I2} are the textual similarities from ASR results, and they play a similar role for the text as Sim_{I1} and Sim_{I2} for the video. The parameter β allows adjusting the relative importance between visual information and textual information.

While in Video-MMR, the basic information unit was a single keyframe, in TV-MMR it is an n-gram segment. The visual content of an n-gram segment is composed of all the keyframes which appear between the beginning and ending times of the utterance. For faster computation, we subsample the video at the rate of 1 frame per second, so that a 5 second utterance will be represented by a set of 5 keyframes. The similarity between keyframes that is used in Video-MMR is extended to a similarity between sets of keyframes by computing the average of keyframes similarities.

The procedure of TV-MMR summarization is explained as the following sequence of steps:

1. The initial video summary S_1 is initialized with one segment, defined as:

   $$S_1 = \underset{f_i, f_i \neq f_j}{\arg\max}[\prod_{j=1}^{n} Sim_I(f_i, f_j) \cdot \prod_{j=1}^{n} Sim_T(f_i, f_j)]^{\frac{1}{n}} \tag{5.36}$$

   where f_i and f_j are n-gram segments from the video set V and n is the total number of segments except f_i. Sim_I computes the similarity of visual information between f_i and f_j; while Sim_T is the similarity of text information between f_i and f_j.

2. Select the segment f_k by TV-MMR formula, Eq. 5.35.

3. Set $S_k = S_{k-1} \bigcup \{f_k\}$.

4. Iterate to step 2 until S has reached the predefined size.

5.4.3 Experimental results of TV-MMR

In the experiments the video sets are still collected from Internet news aggregator website ”wikio.fr”. In this section we totally use 21 video sets in 89 video sets described in Section 3.2.3.1, each of which contains between 3 and 15 videos, whose durations vary from a few seconds to more than 10 minutes. The genres of the videos are various including news, advertisement and movie, to ensure the diversity of the experimental videos.

In the experiment, the similarity of two video frames, $sim(f_i, f_j)$, is defined as cosine similarity of visual word histograms as Eq. 3.2:

$$sim(f_i, f_j) = \cos(H_{f_i}, H_{f_j}) = \frac{H_{f_i} \cdot H_{f_j}}{\|H_{f_i}\| \|H_{f_j}\|} \tag{5.37}$$

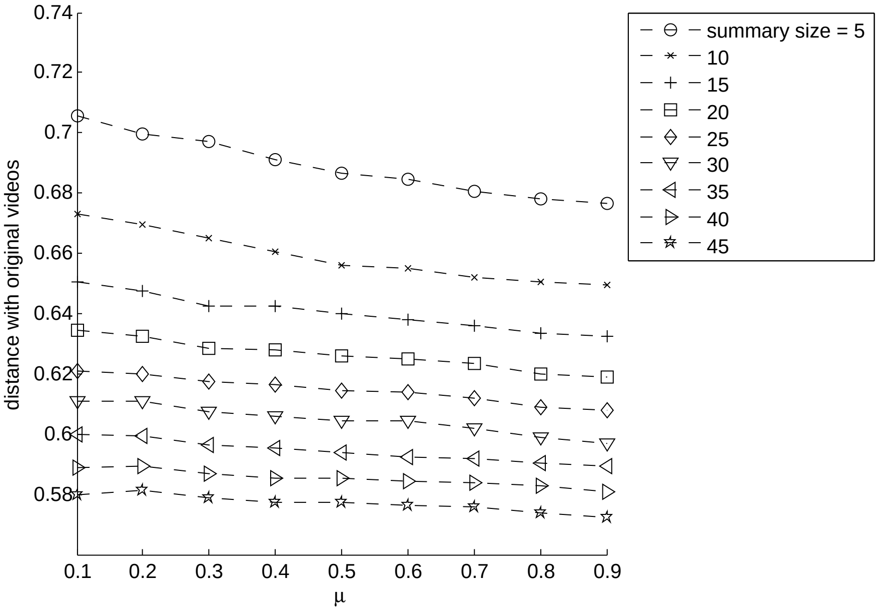


Figure 5.10: SRC of parameter μ

where H_{f_i} and H_{f_j} are histogram vectors of frame f_i and f_j. And for the similarity of text of two segments in TV-MMR, it uses the same definition with Eq. 5.37 but the text histogram of an utterance is defined as:

$$H = (w_1, w_2, \ldots, w_T) \tag{5.38}$$

where w_T is the number of Tst word in the utterance, and the number of the words is T.

To remain consistent with Video-MMR, we still use Summary Reference Comparison (SRC) in Section 3.2.2 to select the best parameters μ and β. First we vary μ from 0.1 to 0.9, each step being 0.1. Then we get a figure for 2-gram as the basic unit in Fig. 5.10.

It is obvious that $\mu = 0.9$ is the best in Figure 5.10. For the other n-grams, the figures are similar with $\mu = 0.9$ owning the best curves, but they are not shown because of the limited pages. Therefore in Eq. 5.35 we prefer $\mu = 0.9$. And we vary β like μ and consider βs for different n-grams, finally we choose $\beta = 0.1$.

Because we have known $\lambda = 0.7$ in Video-MMR, in Eq. 5.35 $\lambda = 0.7$, $\mu = 0.9$ and $\beta = 0.1$. After the best parameters are decided, we can compare the text-visual distances with original videos of TV-MMR and Video-MMR in Fig. 5.11. In Fig. 5.11, we only show the examples of 2-gram and 8-gram, but the other n-grams have similar curves. It is obvious that our TV-MMR outperforms the existing algorithm Video-MMR.

5.5 Static and dynamic summaries

There are two forms of video summary: stationary summary and video skim. In this section we propose two optimization approaches for these two kinds of summaries using both video and text. One approach focuses on static summaries (stationary summary), where the summary is a set of selected keyframes and keywords, to be displayed in a fixed area. The second approach addresses dynamic summaries (video skim) where video segments are selected based on both their visual information and textual content from

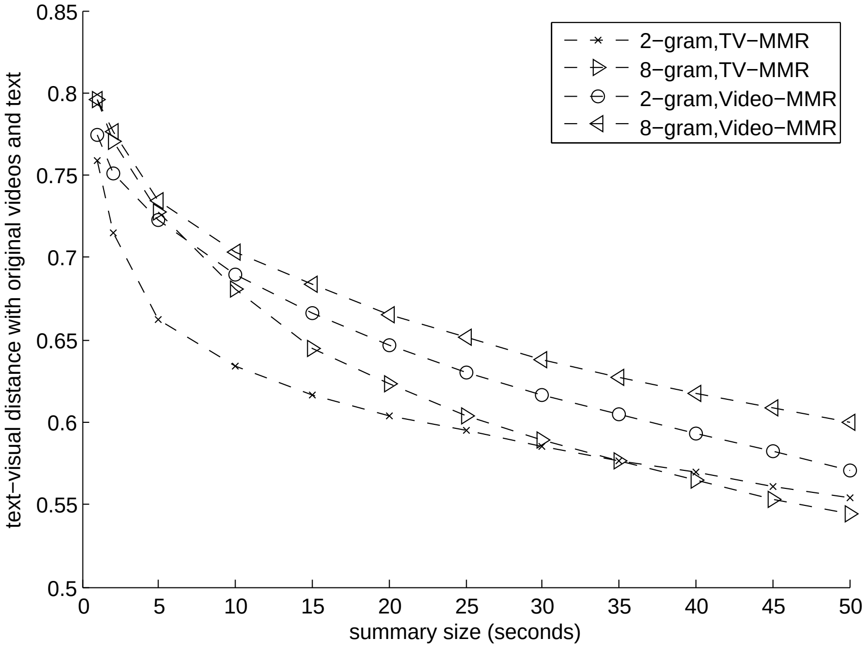


Figure 5.11: TV-MMR and Video-MMR

ASR to compose a new video sequence of predefined duration. The approach for static summary relies on Video-MMR proposed in Section 3.2.2, and the approach for dynamic summary needs to use TV-MMR proposed in Section 5.4.

5.5.1 Static summaries

5.5.1.1 The principle of static summaries

A static video summary is basically composed of selected keyframes. However, it is useful to use also some of the display space to show some keywords which are related to the content of the video sequence. In our work, we use the speech transcription of the audio track, as described in Section 5.4.1. The summary is often presented inside a display space with a predefined size, for example a web page. Therefore, the summarization algorithm has to select a predefined number of keyframes to fit inside this space, while maximizing the amount of information which is presented to the user. When keywords or phrases are also available, the summarization algorithm should decide, not only on which keywords to display, but also about the relative number of keywords and keyframes to fit in the predefined space. The diversity of the visual and the textual content is different from video to video, so that a fixed choice for the number of keywords and keyframes cannot be optimal.

In our work, we have considered that keyframes are of fixed size (another option would be to allow some keyframes to shrink, but we leave it for future exploration), and that the space occupied by a keyframe is equal to the space occupied by 60 characters. Selecting more keyframes reduces the number of words in the space, and vice-versa. For a fixed display space, only combinations of keyframes and keywords which fit inside this space are considered. The task of the summarization algorithm is to find the combination that provides the most information.

Our video summarization algorithm, Video-MMR, is incremental, and produces a sequence of video summaries where one keyframe is added at each step. This provides a sequence of keyframes with decreasing visual importance, out of which we can easily consider the first k, for any value of k. During the Video-MMR, the marginal relevance $k_V(i)$ of a keyframe f_i as defined in Eq. 5.39, decreases as the iterations proceed. We fix a threshold and stop the Video-MMR iterations when the marginal relevance falls below the threshold. For a given video, this provides a number M of keyframes. We normalize the visual relevance of the keyframe:

$$k'_V(i) = k_V(i) / \sum_{j \in M} k_V(j) \tag{5.39}$$

From the speech transcription, we can associate each video keyframe with an n-gram, based on the timecodes. This allows defining the text similarity $k_T(i)$ of the text segment associated to the keyframe f_i as the cosine measure introduced in Section 5.4.1. Again, we normalize these values over the selected set:

$$k'_T(i) = k_T(i) / \sum_{j \in M} k_T(j) \tag{5.40}$$

We take the size of a keyframe as the basic unit, and assume that the available display size is P times the size of a single keyframe. As mentioned previously, the size of a character is taken as 1/60 of the keyframe size. With these figures, the optimal summary will be composed of the set of keyframes ρ_V and the set of keywords ρ_T which satisfy:

- The optimal summary to be presented in a display space, the best combination of frames and text, is the one that maximizes the total visual and textual information that is presented, as is described in the following formula:

  $$\max_{\rho_V, \rho_T} [K_V(\rho_V) + K_T(\rho_T)] \tag{5.41}$$

  With the constraint $size(\rho_V) + size(\rho_T) \leq P$, and the definitions:

  - $K_V = \sum_{j \in \rho_V} k'_V(j)$,
  - $K_T = \sum_{j \in \rho_T} k'_T(j)$,
  - $size(\rho_V) = |\rho_V|$,
  - $size(\rho_T)$ =(number of characters of words in ρ_T)/60,

5.5.1.2 Experimental results of static summaries

The experimental videos of static summaries are the same as TV-MMR in Section 5.4. For the experiments of static summaries, we consider several display sizes:

- $P = 12$, as a reasonable value when the display space is a full screen on a computer,
- $P = 6$, a common value when using the display of a smart phone,
- $P = 3$ and $P = 4$, as often found when a single line of keyframes is considered, inside a larger page.

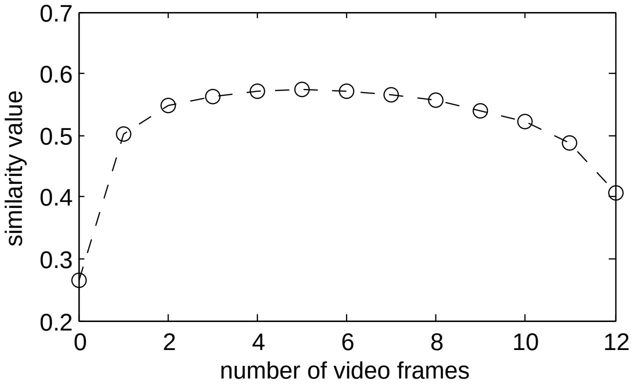


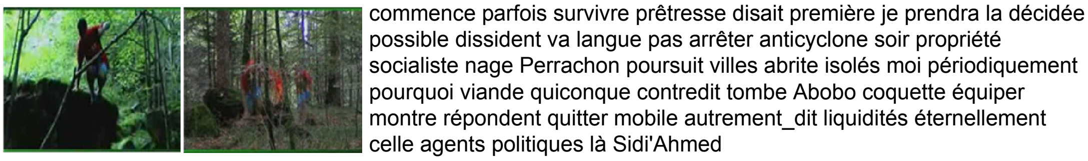
Figure 5.12: Information value of different $|\rho_V|$ when $P = 12$ and $gram_n = 2$

Table 5.12: Statistical data of the best frame number in P

frame	1-gram	2	3	4	5	6	7	8	9	10	sentence
P=12	4	5	5	5	4	4	5	4	5	5	8
P=6	2	2	2	2	2	2	3	2	3	3	4
P=4	1	2	2	2	1	2	1	2	1	2	4
P=3	1	1	1	1	1	1	1	1	2	2	3

We perform experiments over 21 different video sets, representing more than 200 videos. For each set, we consider different values of $|\rho_V|$, select the corresponding keyframes and keywords, and plot the value of the total visual and textual information in the summary, as defined in Eq. 5.41. Figure 5.12 is the curve for the case where the display size is $P = 12$, the text segments are 2-grams, and $|\rho_V|$ varies from 0 to 12. The maximum value is obtained for $|\rho_V| = 5$. Table 5.12 shows the overall results of the optimal value of $|\rho_V|$ for various values of P and various lengths of n-grams. We can see that the optimal number of keyframes has few variations when different lengths of n-grams are considered. However, when full sentences are considered for the text segments, selecting a complete sentences force to select both important and unimportant keywords, which is suboptimal, and only keyframes are selected in the final summary.

In Figure 5.13 we show an example of the static summary for $P = 6$ and 1-gram (For better visualization, the total space is not exactly 6 times the space of an image).



commence parfois survivre prêtresse disait première je prendra la décidée
possible dissident va langue pas arrêter anticyclone soir propriété
socialiste nage Perrachon poursuit villes abrite isolés moi périodiquement
pourquoi viande quiconque contredit tombe Abobo coquette équiper
montre répondent quitter mobile autrement_dit liquidités éternellement
celle agents politiques là Sidi'Ahmed

Figure 5.13: An example of static summary

5.5.2 Dynamic summaries

5.5.2.1 The principle of dynamic summaries

Our dynamic summary is the concatenation of audio-visual segments extracted from the original videos. The candidate segments out of which we select are the segments corresponding to the utterances of n-grams. Segment duration coincides with the utterance of n-gram of text. In this section, we only discuss the dynamic summaries from the viewpoint of maximizing the information in summaries, though the story flow is also important for the dynamic summary.

A specific difficulty comes from the fact that the rate of information flow is different between the audio and the visual channel. For the visual part, videos are a succession of shots. Those shots are often rather long (on the order of 10 seconds or more), with slow motion (with the exception of music clips). In this case, a visual presentation of 1 or 2 seconds of the shot is sufficient to convey most of the visual content of the shot. Any longer presentation is a wasteful usage of the visual information channel for the summary. On the contrary, for the textual part, redundancy is extremely rare, so that longer extracts provide greater information content. Therefore, the choice of the optimal duration of n-grams is lead by two opposite constraints:

- Smaller values of n favor more visual content to be presented (for a given summary duration),
- Higher values of n allow more coherent text information to be included.

Based on this analysis, we explore the use of TV-MMR to find the best compromise between those constraints. For each value of n, we can build a summary from the n-gram segments. We can then compare the quality of these different summaries and select the best one according to a combination of its visual and textual content. We propose the following equation for this optimization:

$$\arg\max_{n} K(S_n) = \arg\max_{n}[K'_V(S_n) + K'_T(S_n)] \tag{5.42}$$

where S_n is the summary built by TV-MMR from the n-gram segments, $K(S_n)$ is the quality of its audio-visual content, defined as the sum of K'_V, the similarity of video segments in the summary with the original video and K'_T, the similarity between text words in the summary and all the text. Before applying TV-MMR, we define the expected duration of the summary.

We then perform experiments to compare the values of text similarities and visual similarities from different values of n, in order to find the best compromise.

5.5.2.2 Example results of dynamic summaries

The experimental videos of static summaries are same as TV-MMR in Section 5.4. To obtain dynamic summaries with the duration D =10, 30, 50, 60, 70, or 80 seconds, we carry out TV-MMR with different grams as the basic unit, 1∼10-gram and sentence. Then we compute text similarities with utterance collection and visual similarities with original videos for dynamic summary as Eq. 5.37 and Eq. 5.38. The mean text similarities and visual similarities of 21 video sets from different n-grams are shown in Fig. 5.14. Each curve in Fig. 5.14 represents textual and visual scores of different n-grams for each summary duration. When the summary duration increases for a fixed n-gram, both textual

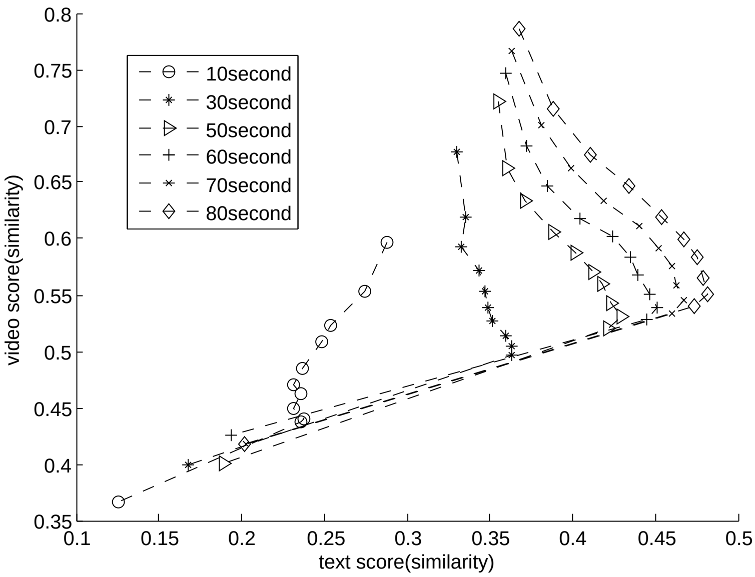


Figure 5.14: Dynamic summaries: the points from top to down are the values of 1∼10-gram and sentences in each curve (the durations of summaries are indicated in the legend)

and visual information increases. When the summary duration is fixed and the number n of n-gram increases, visual information definitely decreases, because less segments represent less visual information and every longer visual segment does not contain much more visual information, but textual information does not monotonously increase because of the compromise of every longer segment representing more meaningful coherent textual information and less number of segments representing less textual grams. When the summary time is short like 10 seconds and 30 seconds, text scores don't increase with the increase of $gram_n$. However, when the summary time is around 60 seconds, $gram_n$ begins to influence on text similarity. The points of 7-gram are inflection and moderate points which maximize both text and video similarities for D =50, 60, 70, or 80 seconds.

Therefore, 7-gram is the best length of the basic unit/segment for dynamic summary, maximizing both text and visual information in a dynamic summary with a duration more than and around 60 seconds. A short basic unit, like 1-gram, seems to be better when the summary size is shorter than 50 seconds. According to our experimental data, the average durations are 2.1 seconds for 7-gram and 0.3 seconds for 1-gram. Therefore in dynamic summary of 60 seconds, every basic segment should last for 2.1 seconds.

5.5.3 Conclusion of static summaries and dynamic summaries

For static summaries we have presented an algorithm which selects keyframes and keywords to maximize the visual and textual information presented in a predefined display space. Our algorithm automatically chooses the optimal number of keyframes. For dynamic summaries based on the concatenation of selected short video segments, we maximize summary information by deciding the best segment duration on average with the help of TV-MMR.

5.6 Conclusion of Multimedia MMR

In this section we extend Video-MMR, which only exploits visual information in the video, to AV-MMR, Balanced AV-MMR and OB-MMR by exploiting both audio and visual information. The experiments have proved that the new algorithms based on both audio and visual information achieve better summaries which are more similar to original videos.

Besides, we extend Video-MMR to TV-MMR by exploiting both textual and visual information in the video. TV-MMR outperforms Video-MMR when computing textual and visual similarity with original videos.

Furthermore, we propose two ways to present and optimize the summary: one is in static form, static summary, and the other one in dynamic form, dynamic summary. We have suggested the best number of frames in the predefined spaces for static summary and the average best duration of the segment for dynamic summary. Both optimized summary forms could present maximum multimedia information to the user.

However, one limit in this chapter is that we separately combine the information of audio and video and the information of text and video. One possible improvement is to combine three indices, text, audio and video together in the future to get a better algorithm of video summarization.

Chapter 6

Video Evaluation by Relevant Threshold

Video Summarization has become an important tool for multimedia information processing, but the automatic evaluation of a video summarization system remains a challenge. People can easily distinguish between ”good” and ”bad” summaries, but an ideal ”best” summary does not exist, so that it is difficult to define a quality measure that can be automatically computed. It is still possible to set up experiments involving human beings to evaluate video summaries, but these experiments are costly, time-consuming, and cannot easily be repeated, which impairs the development of many algorithms based on machine learning techniques. A good quality measure achieving automatic computation, and showing a strong correlation with human evaluation is therefore of great interest.

A similar situation arises in machine translation and text summarization, where specific automatic procedures, respectively BLEU and ROUGE, evaluate the quality of a candidate by comparing its local similarities with several human-generated references. These procedures are now routinely used in various benchmarks. In this chapter, we extend this idea to the video domain and propose the VERT (Video Evaluation by Relevant Threshold) algorithm [Li and Merialdo, 2010c] [Li and Merialdo, 2010d] to automatically evaluate the quality of video summaries. VERT, suitable for both single and multiple videos, mimics the theories of BLEU and ROUGE, and counts the weighted number of overlapping selected units between the computer-generated video summary and several human-made references. Several variants of VERT are suggested and compared. Red, green and blue being the three primary colors, ROUGE, VERT and BLEU, their French translations, could become the set of reference evaluation algorithms in their respective domains too.

6.1 The review of BLEU

Human evaluation of machine translation (MT) needs to consider many factors: adequacy, fidelity, and fluency of the translation [Hovy, 1999]. Although human evaluations are perfect, it is very time consuming, maybe taking weeks or months. Therefore a method of automatic evaluation of MT can help researchers quickly evaluate their MT approaches. For automatically evaluating the quality of machine translation, the BiLingual Evaluation Understudy (BLEU) [Papineni et al., 2002], based on n-gram co-occurrence scoring, has been proposed. BLEU was the scoring metric used in the NIST (NIST 2002) translation benchmarks. The main idea of BLEU is to measure the similarities between a candidate translation and a set of reference translations.

Since a professional human translation is considered as an ideal translation, MT should achieve as close as possible with human translation. It is clear that a better MT shares more words and phrases with the reference translations. Consequently, BLEU should be able to count the number of matches between n-grams in the candidate MTs and n-grams in the reference translations from human. BLEU is a precision method, which calculates the number of n-grams occurring in the references and divides the total number of n-gram in the candidate. The computation is performed sentence by sentence. The results of the BLEU measure have been shown to have a high correlation with human assessments. Precision score P_n is defined as:

$$P_n = \frac{\sum_{C \in \{CandidateSentences\}} \sum_{gram_n \in C} Count_{clip}(gram_n)}{\sum_{C \in \{CandidateSentences\}} \sum_{gram_n \in C} Count(gram_n)} \tag{6.1}$$

where $Count_{clip}(gram_n)$ is the maximum number of n-grams co-occurring in the candidate translation and one of the reference translations, and $Count(gram_n)$ is the number of n-grams in the candidate translation.

Then it is necessary to compute the geometric mean of the precision scores of different gram length - n. The first task is to compute the geometric average of p_n, weighted by w_n. Next, let c denote the length of the candidate translation and r be the effective reference corpus length. A weight called Brevity Penalty (BP) is defined:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} . \tag{6.2}$$

Then

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n). \tag{6.3}$$

Eq. 6.3 is more apparent in the log domain, so the log of Eq. 6.3 is the following

$$logBLEU = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n \tag{6.4}$$

where $N = 4$ and $w_n = 1/N$ in [Papineni et al., 2002].

BLEU can accelerate the development of MT and make researchers quickly decide the performance of their novel algorithms. And in [Papineni et al., 2002] the authors think that BLEU is suitable for evaluating the text summarization because MT and summarization are both natural language generation from a textual context.

6.2 The review of ROUGE

For text summarization evaluation, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) proposed by Lin [Lin and Hovy, 2003][Lin, 2004] automatically determines the quality of a summary compared with human summaries. ROUGE counts the number of overlapping units such as n-gram, word sequences, and word pairs. ROUGE has been used in the Document Understanding Conference (DUC), a large-scale summarization evaluation sponsored by NIST.

In [Lin, 2004], several variants of the measure are introduced: ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S.

1. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. It is defined by the following formula:

$$ROUGE\text{-}N = \frac{\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{6.5}$$

   where n is the length of the n-gram, $gram_n$, Count($gram_n$) is the number of n-grams in the reference summaries, and $Count_{match}$ ($gram_n$) is the maximum number of n-grams co-occurring in a candidate summary and in reference summaries. ROUGE-N is a recall-related measure, while BLEU is a precision-based measure [Lin and Hovy, 2003]. By controlling the grams in the references, we can control the focus on the preference what we want. And since the numerator accumulates the matched n-grams, more weights are given to the n-grams happening in the multiple references. More references a n-gram is included in, the more favorable it is.

2. ROUGE-L prefers the longest common subsequence (LCS) with maximum length between the candidate and the references. If the LCS of two summaries is longer, these two summaries are more similar. In sentence level, ROUGE-L, F_{LCS} here, between the summary X with length m and Y with length n is following:

$$R_{LCS} = \frac{LCS(X,Y)}{m} \tag{6.6}$$

$$P_{LCS} = \frac{LCS(X,Y)}{n} \tag{6.7}$$

$$F_{LCS} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{6.8}$$

   where $LCS(X,Y)$ is the length of a longest common subsequence of X and Y, and $\beta = P_{lcs}/R_{lcs}$. β is set to a very big number in DUC. While for summary-level LCS assume that a reference summary sentence, r_i, contains u sentences with m words and every candidate summary sentence, c_j, contains v sentences with v words, the summary-level LCS-based F-measure can be computed as follows:

$$R_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\bigcup}(r_i, C)}{m} \tag{6.9}$$

$$P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\bigcup}(r_i, C)}{n} \tag{6.10}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{6.11}$$

   In [Lin, 2004] the authors give an example of the summary-level LCS-based F-measure as following: For example, if $r_i = w_1w_2w_3w_4w_5$, and C contains two sentences: $c_1 = w_1w_2w_6w_7w_8$ and $c_2 = w_1w_3w_8w_9w_5$, then the longest common subsequence of r_i and c_1 is "w_1 w_2" and the longest common subsequence of r_i and c_2 is "$w_1w_3w_5$". The union longest common subsequence of r_i, c_1, and c_2 is "$w_1w_2w_3w_5$" and $LCS_{\bigcup}(r_i, C) = 4/5$.

3. ROUGE-W prefers the weighted longest common subsequence more than ROUGE-L, because ROUGE-L does not consider different spatial relations of different LCSes. Weighted LCS (WLCS) is improved based on LCS by remembering the length of consecutive matches of two summaries. And the definition of ROUGE-W is:

$$R_{wlcs} = f^{-1}(\frac{WLCS(X,Y)}{f(m)}) \tag{6.12}$$

$$P_{wlcs} = f^{-1}(\frac{WLCS(X,Y)}{f(n)}) \tag{6.13}$$

$$F_{wlcs} = \frac{(1+\beta^2)R_{wlcs}P_{wlcs}}{R_{wlcs}+\beta^2 P_{wlcs}} \tag{6.14}$$

   where f^{-1} is the inverse function of f. And the definitions of X, Y, m, and n are the same with Eq. 6.8.

4. ROUGE-S is the Skip-Bigram Co-occurrence Statistics, where a skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. For example, in the following example [Lin, 2004]:

   - S1. police killed the gunman
   - S2. police kill the gunman
   - S3. the gunman kill police
   - S4. the gunman police killed

   S1 has 6 skip-bigrams, defined as $C(4,2) = 6$: ”police killed”, ”police the”, ”police gunman”, ”killed the”, ”killed gunman”, and ”the gunman”. Then S2 has three skip-bigram matches with S1 (”police the”, ”police gunman”, and ”the gunman”), S3 has one skip-bigram match with S1 (”the gunman”), and S4 has two skip-bigram matches with S1 (”police killed”, and ”the gunman”).

   Given the reference translation X with length m and a candidate translation Y with length n, ROUGE-S is formulated as

$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2}+\beta^2 P_{skip2}} \tag{6.15}$$

   and

$$R_{skip2} = \frac{SKIP2(X,Y)}{c(m,2)} \tag{6.16}$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{c(n,2)} \tag{6.17}$$

   where SKIP2(X,Y) is the number of skip-bigram matches between X and Y, β controls the relative importance of P_{skip2} and R_{skip2}, and C is the combination function.

6.3 VERT principle

By borrowing ideas from ROUGE and BLEU, we extend these measures to the domain of video summarization. We focus our approach on the selection of relevant keyframes, as a video skim can be easily constructed by concatenating video clips extracted around the selected keyframes. Since we believe that the temporal order of keyframes is not as important as the word order in a sentence, we rather use the keyframe importance rank in the selection. The process of video summarization is then formalized as follows:

1. We consider a set of video sequences V_1, V_2,…, V_k related to a given topic,
2. These sequences are segmented into shots or subshots, and each shot is represented by one or more keyframes,
3. Based on shots, subshots or keyframes, a selection of the video content to be included in the summary is performed. Eventually, this selection may be ordered, with the most important content being selected first.
4. The selected content is assembled into a video summary, either in the form of a photo album or a video skim.

After the selection, each keyframe is assigned an importance weight $w_S(f)$ depending on the rank of keyframe f in the selection S. Therefore, our VERT measure compares a set of computer-selected keyframes with several reference sets of human-selected keyframes. Since BLEU is a precision measure and ROUGE is a recall measure, we propose VERT-Precision (VERT-P) and VERT-Recall (VERT-R) respectively.

6.3.1 VERT-Precision

Mimicking BLEU algorithm, we propose the VERT-Precision (VERT-P) measure. Assume that we have k reference summaries (human selected lists), each containing n keyframes. Each keyframe is assigned an importance weight $W_S(x, y)$ according to its position in the selection ($x = 1, \dots, k$ and $y = 1, \dots, n$). Non-selected keyframes are assigned a weight of zero. Similarly the candidate summary (computer selected list) contains m keyframes, and each keyframe i is assigned a weight $W_C(i)$. VERT-P measures the precision of the position of each candidate keyframe in comparison to the reference summaries. For each keyframe i in candidate summary, the maximum weight that is assigned in the reference summaries is

$$T_i = \max_x W_S(x, y_i) \tag{6.18}$$

where x is a reference summary, and y_i is the position of keyframe i in x. VERT-P compares this maximum weight with the actual weight that keyframe i is assigned in the candidate summary. The definition of VERT-P is:

$$\mathit{VERT\text{-}P} = \frac{\sum_{i=1}^{m} \min[W_C(i), T_i]}{\sum_{i=1}^{m} W_C(i)} \tag{6.19}$$

The value of VERT-P is always between zero and one. The maximum is obtained when every keyframe of the candidate was selected with a weight that is lower than at least one of the human selections. For example, if a candidate keyframe was never selected by any human, the value of the measure will be strictly lower than 1. In this way VERT-P is a precision-based measure.

6.3.2 VERT-Recall

By similarity with ROUGE-N, we propose VERT-R$_{\mathrm{N}}$:

$$\mathit{VERT}\text{-}R_N(C) = \frac{\sum_{s\in\{ReferenceSummaries\}}\sum_{gram_n\in S} W_C(gram_n)}{\sum_{s\in\{ReferenceSummaries\}}\sum_{gram_n\in S} W_S(gram_n)} \tag{6.20}$$

where C is the candidate video summary, $gram_n$ is a group of n keyframes, $W_S(gram_n)$ is the weight of the group $gram_n$ for a reference summary S, and $W_C(gram_n)$ is the weight of the group $gram_n$ for the candidate summary C. Note that in the numerator of the formula, the summation of $W_C(gram_n)$ is only taken for the $gram_n$ which are present in the reference summary S.

VERT-R$_{\mathrm{N}}$ is a recall-related measure too. As VERT-R$_{\mathrm{N}}$, it computes a percentage of $gram_n$ from the reference summaries occurring also in the candidate summary. While ROUGE uses the notion of ”word matching”, VERT-R considers the notion of ”keyframe similarity”, which may be interpreted in a very strict sense (selection of the same keyframe), but also in a more relaxed manner by introducing a similarity measure between keyframes.

When n is larger than 1, the notion of ”group of n keyframes” may have several interpretations. Since the selected summaries are ranked lists of keyframes, it is possible to consider consecutive keyframes in these lists. However, we decided that it was more sensible to define a ”group of n keyframes” as a simple subset of size n, because the proximity of keyframes in the selected lists does not bear as much information as the order of words in a sentence. In this regard, VERT-R$_{\mathrm{N}}$ resembles more to ROUGE-S.

In this chapter, we restrict our study to the cases $n = 1$ and $n = 2$. We thus define VERT-R$_1$ and VERT-R$_2$ measures by Eq. 6.21:

$$\begin{aligned} \mathit{VERT}\text{-}R_1(C) &= \frac{\sum_{S\in R}\sum_{f\in S} W_C(f)}{\sum_{S\in R}\sum_{f\in S} W_S(f)} \\ \mathit{VERT}\text{-}R_2(C) &= \frac{\sum_{S\in R}\sum_{(f,g)\in S} W_C(f,g)}{\sum_{S\in R}\sum_{(f,g)\in S} W_S(f,g)} \end{aligned} \tag{6.21}$$

In VERT-R$_1$, each $gram_1$ contains only 1 keyframe, so that the number of $gram_1$ is just the number of keyframes, and the weight of a group is simply the weight of the keyframe. Note that the denominator in Eq. 6.21 is actually the product of the total number of keyframes in all reference summaries times the sum of all weights. It’s a one-dimension computation.

In VERT-R$_2$, there are 2 keyframes in each $gram_2$, so it requires a two-dimension computation. We propose two variants for VERT-R$_2$:

1. VERT-R$_{\mathrm{2S}}$, where the weight of a $gram_2$ is the average of the weights of the keyframes:
$$W_S(f,g) = \frac{w_S(f) + w_S(g)}{2} \tag{6.22}$$

2. VERT-R$_{\mathrm{2D}}$, where the weight of a $gram_2$ is the absolute difference between the weights:
$$W_S(f,g) = |w_S(f) - w_S(g)| \tag{6.23}$$

Obviously, VERT-R$_{\mathrm{2D}}$ should only be considered if weights are non-uniform.

6.4 Experimental results of VERT

For our experiments of VERT, we use two sets of videos, ”DATI” and ”YSL” same as Section 3.2.2. We use the video summarization algorithm Video-MMR from Section 3.2.2 for the initial keyframe representation of the videos.

This section is organized as follows: Subsection 6.4.1 explains the method for constructing the references by human assessment, and two systems of weights are suggested: ranking weights and uniform weights; Subsection 6.4.2 explains the principle of VERT evaluation; and the evaluation results are presented in Subsection 6.4.3.

6.4.1 Reference construction

We now detail how we organized the construction of human-selected summaries which would be used as references. Our concern is to design a process which would facilitate the selection as much as possible, despite the complexity of the task.

1. For each video set, we identify 6 representative videos. For this, we compute the mean distance between each video and all the others in the set. Then we select the 3 videos with the smallest means, as containing the core of the set, and the 3 videos with the highest means, as containing the most distinctive information from the set.

2. On these 6 videos, we perform shot boundary detection, and one representative keyframe per shot is selected.

3. If a video produces more than 10 keyframes, we select 10 most important keyframes by the Video-MMR algorithm. The result is a set of at most 60 keyframes that is representative of the visual content of the video set.

4. From these 60 keyframes, we ask each user to select 10 most important frames as reference summaries. The selection is ordered, with the most important frame being selected first. Users may watch the original video if desired, and they can also access the related textual information.

The summaries of video sets ”DATI” and ”YSL” for constructing the references are shown in Fig. 6.1 and Fig. 6.2. The images in the same row originate from the same video. We enrolled 12 users, members of other projects in the laboratory, to select their own best summaries of 10 keyframes. In the Ranking Weights scheme, the weights decrease linearly from 1.0 (for the most important frame) to 0.1 (for the least important). In the Uniform Weights scheme, the weights are all equal to 0.1.

6.4.2 VERT evaluation principle

We want to evaluate if the values assigned by the VERT measures correlate with the human judgment on the quality of summaries. For each set ”DATI” and ”YSL”, we constructed a set of 7 representative summaries of 10 keyframes each, including 2 random summaries, 1 summary constructed by K-Means, 2 summaries constructed by Video-MMR (with different parameter values), and the best and worst human summaries (based on our own judgment). From these 7 summaries, we created 21 pairs (an example from ”DATI” is shown in Fig. 6.3 and an example from ”YSL” is in Fig. 6.4) and asked humans to select the best summary in each pair. To reduce the load on users, the 21 pairs were separated

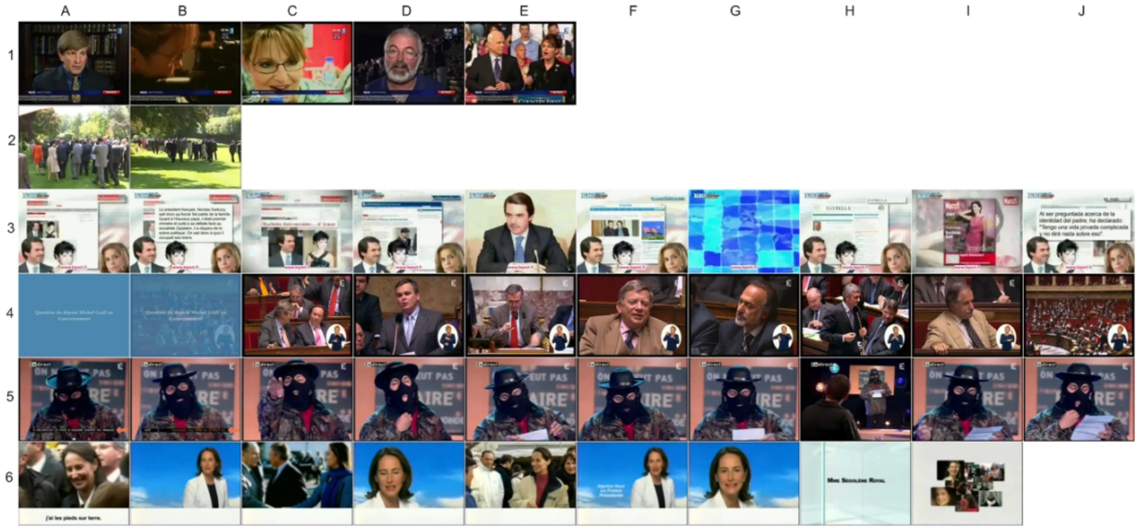


Figure 6.1: The set of keyframes of ”DATI”

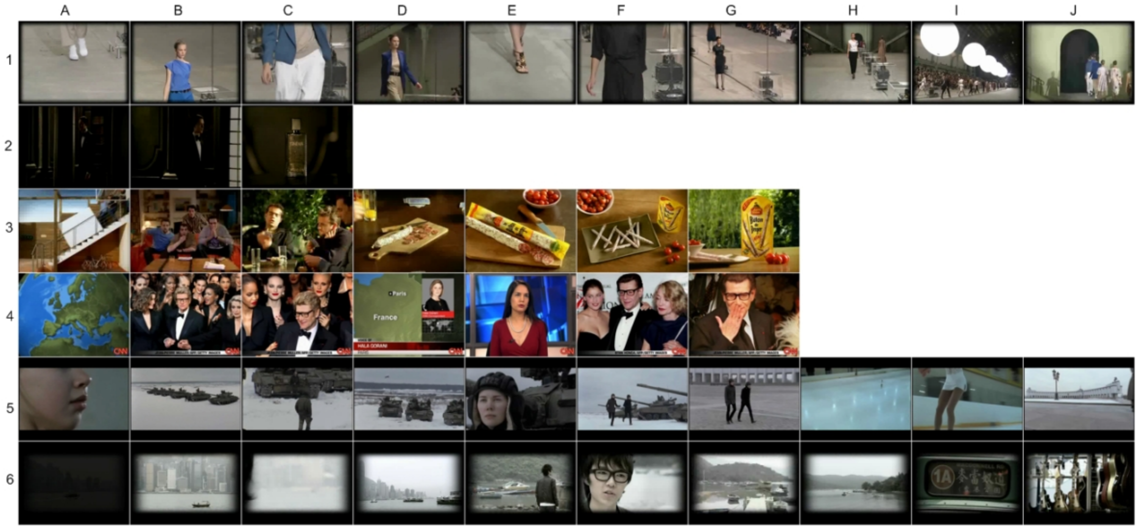


Figure 6.2: The set of keyframes of ”YSL”

Figure 6.3: A Summary Pair of "DATI": One row, one summary.

Figure 6.4: A Summary Pair of "YSL": One row, one summary

into 2 groups and each group was evaluated by 6 users. In total, each pair has 6 evaluations (Human Pair Selection, HPS) from different humans identifying the best summary in the pair.

6.4.3 VERT evaluation

By applying VERT to the same 21 pairs, we can define the VERT Pair Selection (VPS), and compare it with the human selection (HPS). We use the accuracy percentage λ, and the Spearman rank correlation coefficient ρ [Lin and Hovy, 2003][Lin, 2004] to quantify their correlation.

The Accuracy Percentage (AP) is the percentage of correct choices made by VPS compared with human reference, HPS. Let $P_i, i = 1, , \dots 21$ be the 21 pairs, we define:

$$C_{VERT}(i) = \begin{cases} -1, & \text{if the first summary is selected;} \\ +1, & \text{if the second summary is selected.} \end{cases} \tag{6.24}$$

with a similar definition $C_{H_m}(i)$ for the choices of human m. The AP is defined as:

$$\lambda = \frac{1}{H}\sum_{m=1}^{H}[\frac{1}{21}\sum_{i=1}^{21}\frac{C_{VERT}(i)\cdot C_{H_m}(i)+1}{2}] \tag{6.25}$$

where $H = 6$ here. For the Spearman coefficient, we derive a ranking of the 7 summaries from VPS and HPS, and apply the formula:

$$\rho = 1 - \frac{1}{H}\sum_{m=1}^{H}\frac{6}{21(21^2-1)}[\sum_{i=1}^{21}(rank_{VERT}(i) - rank_{H_m}(i))^2] \tag{6.26}$$

Table 6.1 and Table 6.2 show the APs and Spearman coefficients of VERT-P, VERT-R$_1$, VERT-R$_{2S}$ and VERT-R$_{2D}$ for the Ranking Weights system. We also evaluate human selection as the average of each HPS with the other 5 as references. We see that the VERT-R results are in the same range as human evaluation. For Uniform Weights, APs and Spearman coefficients are shown in Table 6.3, which does not contain VERT-P, because VERT-P is meaningless for Uniform Weights.

Table 6.1: λs with Ranking Weights

	P	R_1	R_{2S}	R_{2d}	User
DATI	0.5317	0.6270	0.5794	0.6270	0.5714
YSL	0.5317	0.7063	0.6905	0.6587	0.6286

Table 6.2: ρs with Ranking Weights

	P	R_1	R_{2S}	R_{2d}	$User$
DATI	0.1071	0.6429	0.4643	0.6429	0.6190
YSL	0.2143	0.7500	0.8571	0.8214	0.6310

Table 6.3: λs and ρs with Uniform Weights

	$\lambda(R_1)$	$\lambda(R_{2S})$	$\rho(R_1)$	$\rho(R_{2S})$
DATI	0.6270	0.5794	0.6429	0.4643
YSL	0.6905	0.6905	0.6071	0.8214

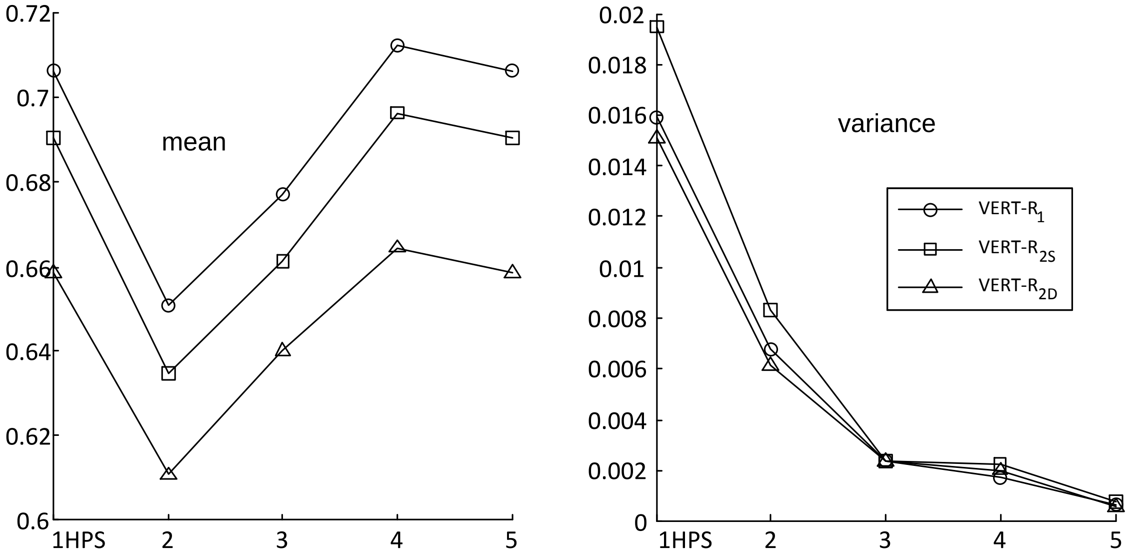


Figure 6.5: Means and Variances of λ for "YSL"

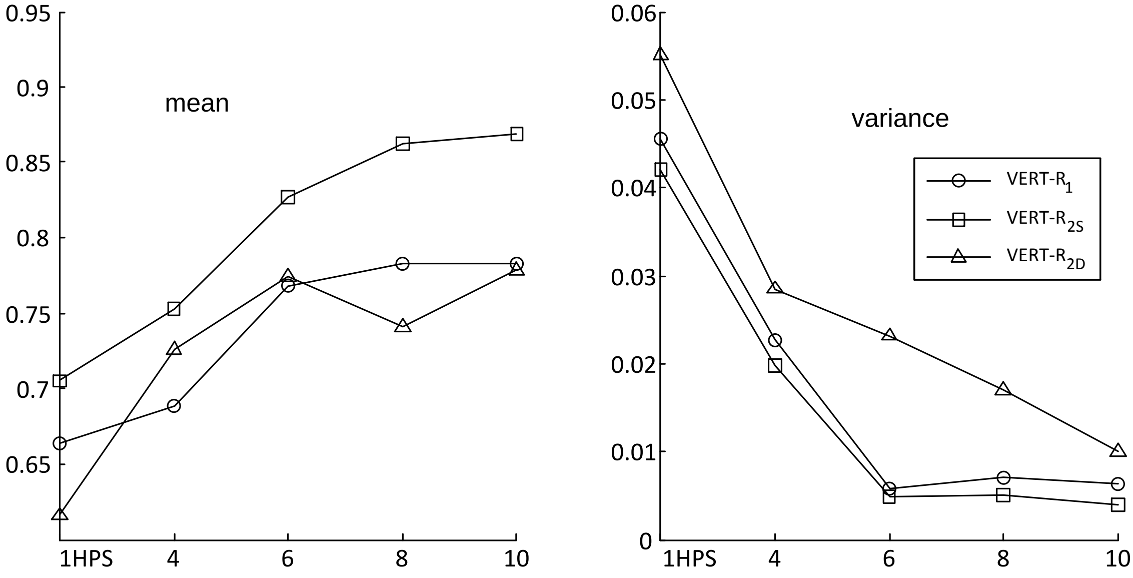


Figure 6.6: Means and Variances of ρ for "YSL"

In Fig. 6.5 and Fig. 6.6, we vary the number of HPS that are considered in the evaluation, from 1 to 5. We see that there is a convergence of the mean values λ and ρ, and that the variances of these values are greatly reduced when the full reference set is used. This is a clear indication that the values that have been computed are reliable estimates.

It is clear that the values of Spearman coefficients for VERT-P are very small, which indicates that VERT-P does not match well with human assessment. For VERT-R, the value of APs and Spearman coefficients are both around 0.6. Since these results are in the same range of value as those presented in [Lin and Hovy, 2003], we can conclude that the VERT-R measure is effective in the summary evaluation.

Through above experiments, VERT-Precision variant has not been found to be effective, while VERT-Recall variant has shown a good correlation with human assessment. With VERT-Recall, several variants have similar performance, so it is hard to choose the best variant. In the future we plan to extend our experiments in size and scope to further identify the capabilities and limitations of the method.

6.5 The demonstration tool of VERT

In this section we present a demonstration tool to visualize the evaluation procedure of VERT. We design an interface for the user to conveniently watch the videos, and select his candidate summaries. Then in the demo tool, the user selection is compared with the reference by the colored grids, and the VERT weights belonging to user selection and the reference are illustrated by two colored grids too. Finally, VERT scores are shown as the curves in the demo tool.

Since we have proved that VERT-Recall is better than VERT-Precision in Section 6.4, in this section VERT only refers to VERT-Recall. For videos used in this section, we use 6 videos of "YSL" video set about the death of a famous designer, which is the same with Fig. 6.2 in Section 6.4. Some videos represent the burial, some display older fashions shows,

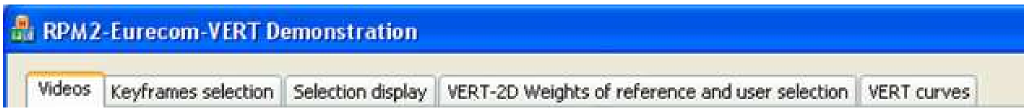

RPM2-Eurecom-VERT Demonstration
Videos
Keyframes selection
Selection display
VERT-2D Weights of reference and user selection
VERT curves

Figure 6.7: Tabs of VERT Demonstration

while some are interviews or comments. It is possible that some videos are incorrectly classified and unrelated to the topic.

Our demonstration tool [Li and Merialdo, 2010a] is designed as a tabbed dialog box. In this dialog, there are five tabs: the first one is ”6 videos”; the second one is ”keyframes selection”; the third one is ”Selection show”; then ”VERT-2D”, (”VERT-R_{2D}” in Section 6.3) weights of reference and user selection” tab is shown; the name of the last tab is ”VERT curves”. These tabs are shown in Fig. 6.7. We will describe these five tabs in detail in the rest part of this section.

6.5.1 The tab ”6 Videos”

In this tab, 6 videos in our prepared video set ”YSL” are played in the demonstration tool.

1. ”Video 1” is a long video of 7 minutes 38 seconds, whose content is the fashion show of the female clothes.

2. ”Video 2” is a video with the length of 29 seconds. It is an advertisement of the perfume.

3. ”Video 3” is to advertise a kind of sausage. Its time length is 19 seconds.

4. ”Video 4” with the length of 2 minutes 20 seconds is the news of CNN about a French clothes designer.

5. ”Video 5” is a MV, which tell us something of skating, arms and so on. Video 5 owns the time of 3 minutes 12 seconds.

6. The last video ”Video 6” is like an interview of a singer from Hong Kong by the brand ”Cartier”. The length of the time is 3 minutes 6 seconds. The design of this tab ”6 Videos” is shown in the following figure, Fig. 6.8.

6.5.2 The tab ”keyframes selection”

This tab is composed of 47 keyframes with the shape of 6 rows and 10 columns. The keyframes are obtained by Video-MMR explained in the Section 3.2.2. Each row is from a video, and the order of columns is as the time order. In Fig. 6.9, there is a check box under each keyframe. The user could select 10 keyframes by the check boxes and at last click the button ”Finish selection”. The reason of selecting 10 keyframes is that we own a reference set of keyframes with the size 10 keyframes by 12 people. After the selection of keyframes, the demonstration tool will compute the demo images and curves explained in Subsection 6.5.3, 6.5.4 and 6.5.5.

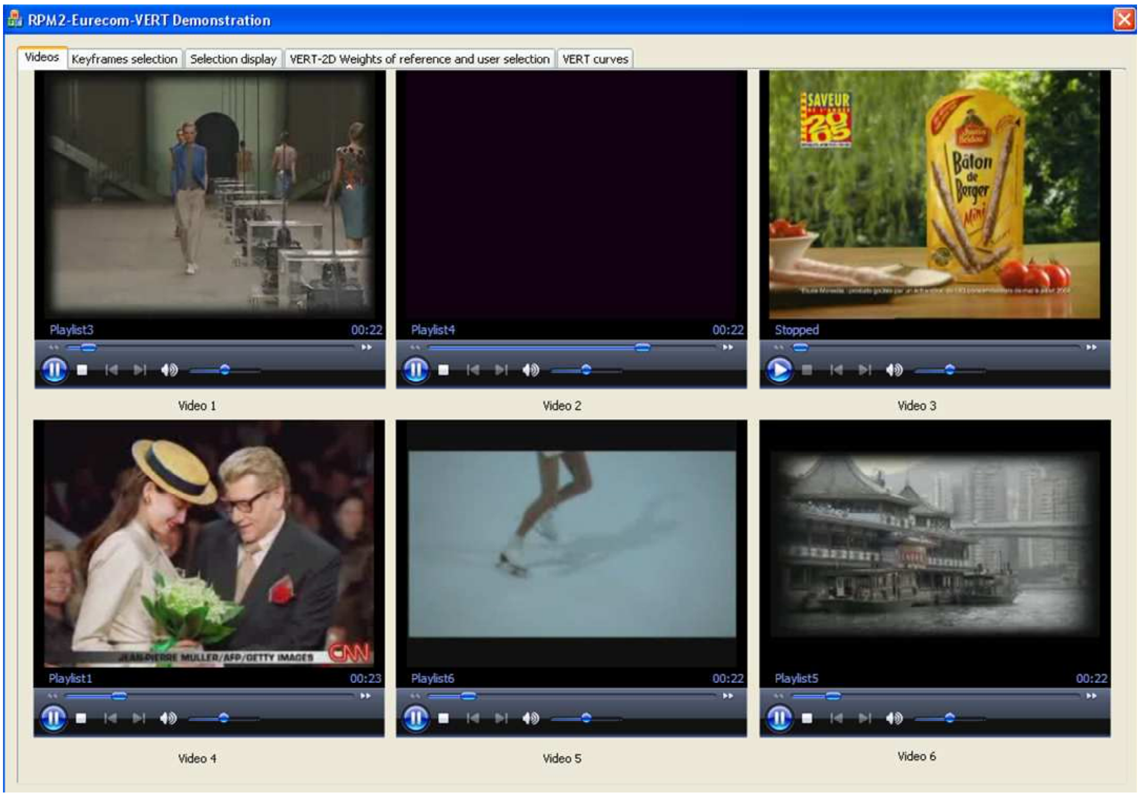
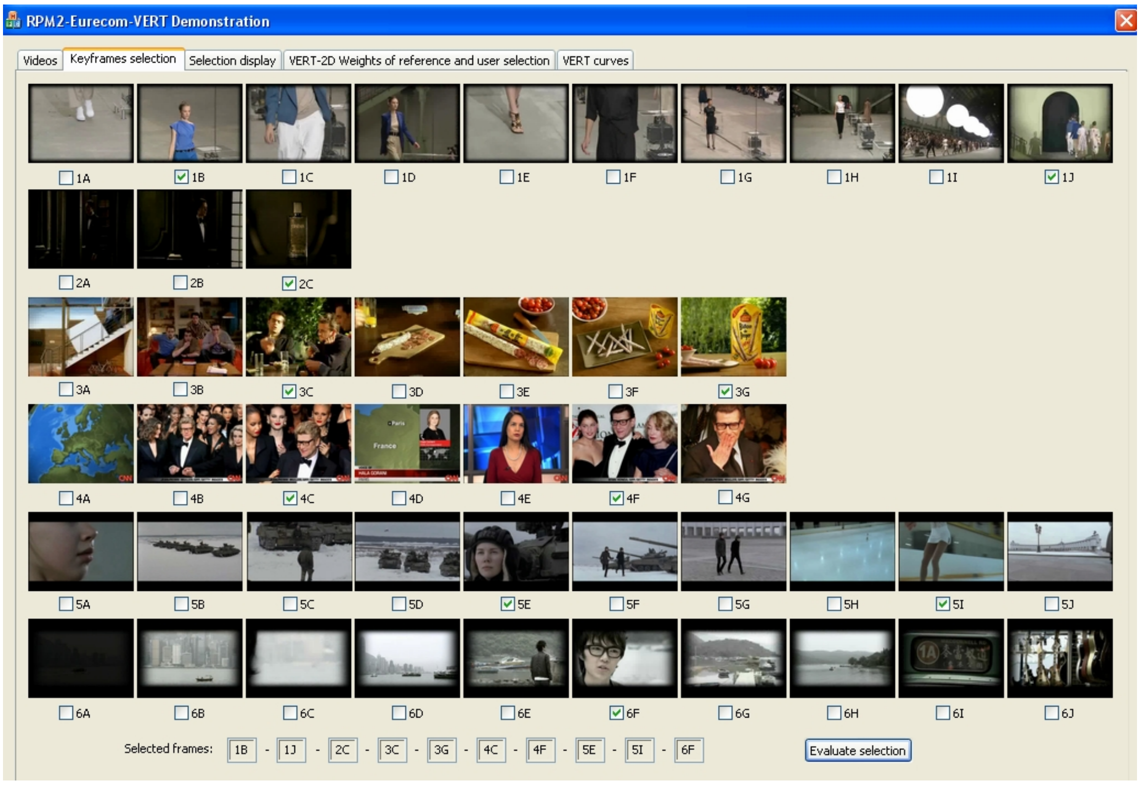


Figure 6.8: The tab "6 Videos"


RPM2-Eurecom-VERT Demonstration
Videos
Keyframes selection
Selection display
VERT-2D Weights of reference and user selection
VERT curves
Evaluate selection

□ 1A ☑ 1B □ 1C □ 1D □ 1E □ 1F □ 1G □ 1H □ 1I ☑ 1J

□ 2A □ 2B ☑ 2C

□ 3A □ 3B ☑ 3C □ 3D □ 3E □ 3F ☑ 3G

□ 4A □ 4B ☑ 4C □ 4D □ 4E ☑ 4F □ 4G

□ 5A □ 5B □ 5C □ 5D ☑ 5E □ 5F □ 5G □ 5H ☑ 5I □ 5J

□ 6A □ 6B □ 6C □ 6D □ 6E ☑ 6F □ 6G □ 6H □ 6I □ 6J

Selected frames: 1B - 1J - 2C - 3C - 3G - 4C - 4F - 5E - 5I - 6F

Figure 6.9: The tab "keyframes selection"

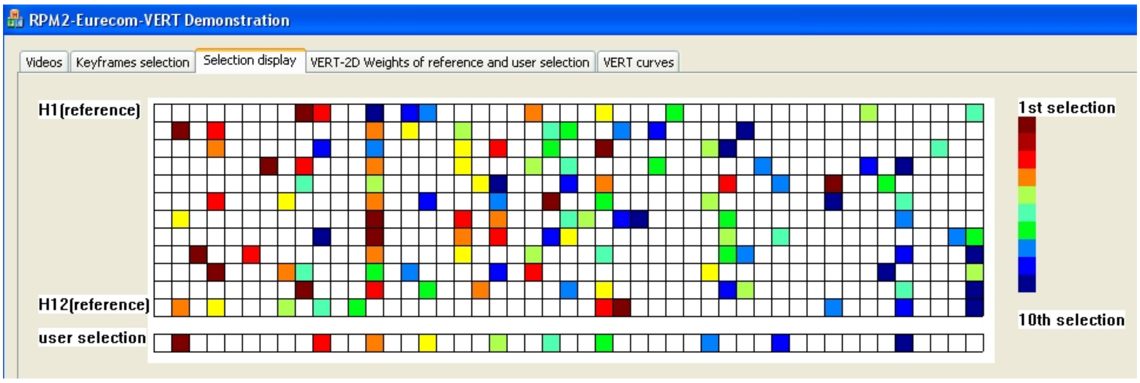


Figure 6.10: The tab ”selection show”

6.5.3 The tab ”selection show”

When the user clicks the tab ”selection show”, an image representing the selected keyframes by reference and user are shown in a grid of 13*47. The number 47 means the total 47 keyframes whose order is from the first video to the sixth video. Furthermore, the first row to the twelfth row is the selection by 12 people in the reference. The last row in the grid is just the user selection in the tab ”keyframes selection”. The selection order is represented by the color. The color bar is shown right to the image in Fig. 6.10. From this Fig. 6.10, we can intuitively read the user selection and reference selections of the keyframes.

6.5.4 The tab ”VERT-2D weights of reference and user selection”

This tab contains two images of 47*47 grids. The number 47 owns the same meaning with the last subsection. Because VERT-2D is 2-D computation and it considers the time order of two keyframes in a 2-gram, the grids are two dimensions. The left grid presents the reference weights of VERT-2D which are the denominator of Eq. 6.23, while the right grid is the user weights of VERT-2D being the numerator of Eq. 6.21. The color bar shows the color from high weight to low weight in Fig. 6.11 too. We can directly view which weights of VERT-2D are selected by the user.

6.5.5 The tab ”VERT curves”

Fig. 6.12 shows two VERT-2D curves, with 12 discrete points. The point ”U, H1” means that it uses the 11 reference summaries excluding the first one as the new reference in Eq. 6.21, the first reference summary H1 as one candidate summary and user selection as another candidate summary. Green point of ”U, H1” displays the VERT-2D score of candidate summary H1 and Red point is the VERT-2D score of candidate summary of user selection. The other points in Fig. 6.12 are similar. We can see that two curves are coherent, which means the reference is stable and trustable.

6.5.6 Conclusion of VERT demonstration tool

Based on the principle of VERT, we designed our demo tool to facilitate the procedure of user selection and the visualization of summary evaluation. With our demo tool, the user could easily analyze the similarity and difference between his selection of video keyframes and the reference. When the selection from the user is more similar to the reference, the

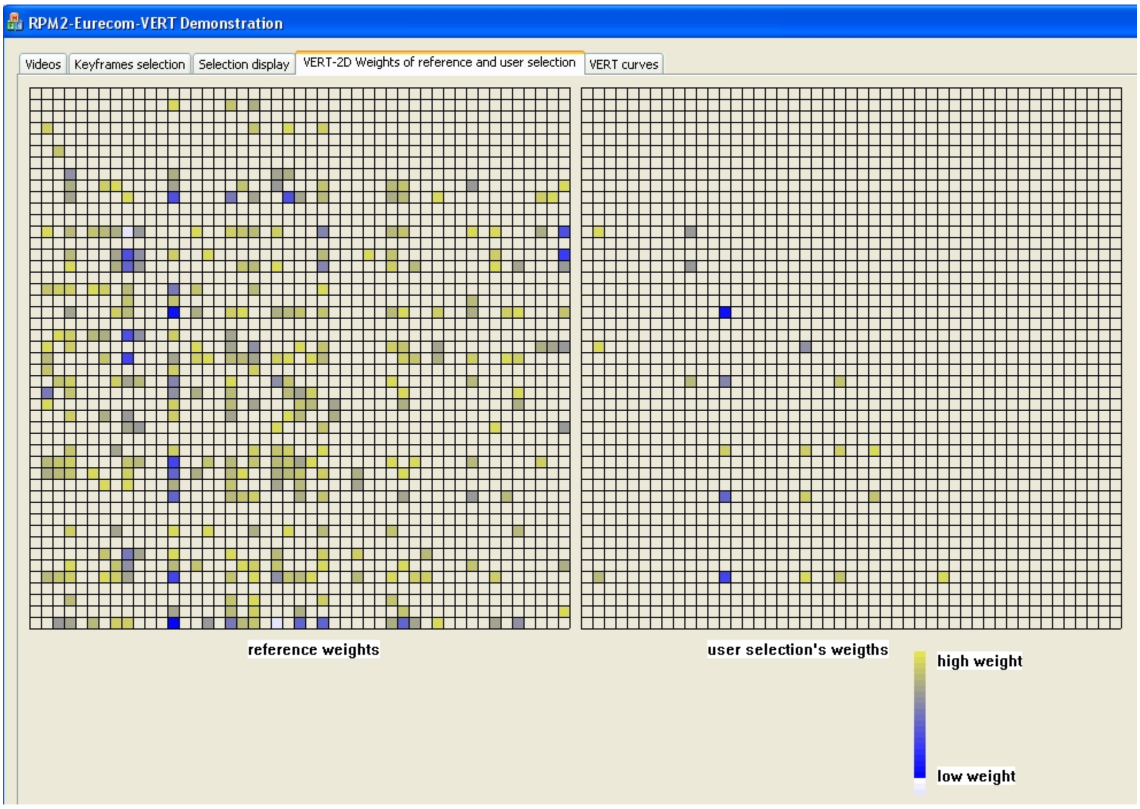


Figure 6.11: The tab ”VERT-2D weights of reference and user selection”

corresponding curves are more coherent. Therefore, our VERT demonstration tool is a good method to visualize the evaluation of video summaries.

6.6 Conclusion of VERT

In this chapter we have proposed an automatic evaluation measure for video summary, VERT, which facilitates the evaluation procedure and avoids the limits of human assessment. Furthermore, we also design a simple demonstration tool for VERT to visually illustrate and demonstrate the procedure of VERT.

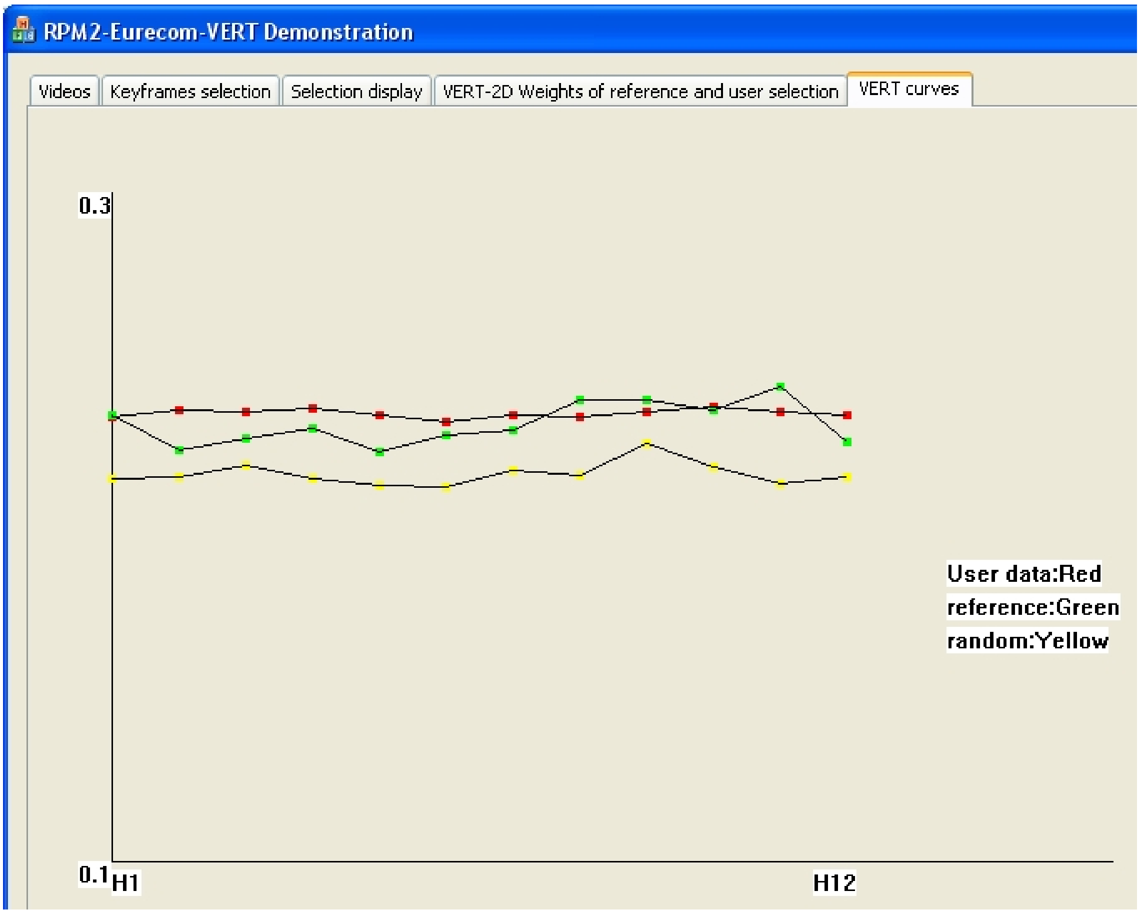


Figure 6.12: The tab ”VERT curves”

Chapter 7

Conclusion

The quantity of the video tremendously becomes more and more every day, so the number of the video from Internet, personal DV, TV, and other sources is out of the management of human being. Video summarization is an important measure to deal with the video by extracting the important information from the video. Video summarization is suitable for interactive browsing and searching systems, by which the user can easily manage and access video content.

In this thesis we have proposed a summarization algorithm, Video-MMR by exploiting the visual information for multi-video. Video-MMR borrows the idea from the algorithm of text summarization, MMR. Video-MMR is an incremental algorithm to construct the summary frame by frame.

Then we develop Video-MMR in two ways:

- One way is to improve Video-MMR itself by exploring the advantages and avoiding the limits of Video-MMR. Consequently we proposed an improved algorithm, Video-MMR2 which only exploits the visual information too. Video-MMR2 avoids selecting the visually unimportant frames for the stationary video summary, including the frame with a major color or the dull frame. The most important improvement of Video-MMR2 is that it avoids selecting duplicated frames into the summary.

- We exploit the information from the audio or the text of the video. By introducing the audio information, we have proposed the algorithm, AV-MMR, Balanced AV-MMR and OB-MMR step by step. Since audio information is important in the video, we obtain better algorithms than Video-MMR. At the meantime, we have introduced the text transcribed from the speech into Video-MMR to form TV-MMR. In addition, for two forms of the summary - static summary and dynamic summary, we have separately suggested the optimized number of keyframes and text grams in a predefined stationary space by Video-MMR and the optimized segment duration of the dynamic summary in video form by TV-MMR.

The video summarization algorithm is only one facet in the domain of video summarization, because the evaluation algorithm, especially the automatic evaluation algorithm, is the other important facet. Without a good evaluation algorithm, we cannot decide the performance of our novel summarization algorithm. While human assessment is an ideal evaluation method, but it is time-consuming, unrepeatable, and unstable. So an automatic evaluation is necessary for the domain of video summarization. Consequently, we have proposed VERT to automatically evaluate the quality of the stationary summary.

With VERT we just need a few reference summaries from human being, then we can automatically get the quantitative qualities of different summaries produced by our algorithms. VERT is time-saving, repeatable and stable compared to human assessment.

7.1 Perspective

Though many research works have been devoted into video summarization, there are still some challenges and possible improvements in several aspects [Money and Agius, 2007], which are suitable for our work too:

1. Exploiting semantic features. For many current algorithms in the domain of video summarization, including the algorithms proposed in this thesis, are lack of exploiting semantic features. So these summarization cannot overcome the semantic gap. Furthermore, most current algorithms cannot produce personalized video summaries.

2. Exploiting external and hybrid features. Exploiting the user-based information outside the video itself can facilitate the production of the personalized summary, for example using the information of the user profiles and the relations of friends in social media.

3. Building and standardizing the evaluation of video summaries. Most current evaluation approaches are quiz methods answered by human. And there is not any successful automatic evaluation measure of video summaries, then it is difficult to compare different algorithms of video summarization. Even in this thesis we have proposed VERT measure, but it only can work for one summary form, storyboard. The wonderful situation for the evaluation is to build a standardized automatic evaluation measure in the feature like the evaluation in TRECVID.

Refer to the work in this thesis, in the future we can improve our summarization algorithms and evaluation measure in the following aspects:

1. We can introduce audio and the text information together into Video-MMR2 by mimicking OB-MMR, which may result in a better algorithm than OB-MMR.

2. Currently the optimized set of parameters in OB-MMR is generic, but OB-MMR is possible to benefit from genre-based parameters for the videos of different genres. The same situation may happen in other proposed algorithms.

3. In OB-MMR we only use some simple semantic features, but it is possible to exploit some high-level semantic information in the video, for example specific objects and highlights.

4. VERT is only suitable for stationary summary composed of keyframes, so VERT cannot deal with the summary with the audio or the text. It is possible to extend the application scope of VERT by considering the audio or the text too.

5. It is meaningful to exploit a crowdsourcing platform such as Amazon's Mechanical Turk to obtain abundant ground truth from human beings to refine the weights in different MMR methods and the selection of VERT.

Appendix A

Résumé (French summary)

L'augmentation rapide de la quantité des vidéos sur Internet est maintenant évidente. Chaque jour, des milliers de personnes téléchargent et partagent de nouvelles vidéos, enreistrements télé, clips musicaux, vidéos personnelles etc... Comment gérer une telle quantité de données visuelles est un problème pour les utilisateurs, et c'est donc un sujet de recherche extrêmement actif de nos jours. Le résumé vidéo est reconnu comme un élément important dans le traitement des données vidéo. Un système de résumé vidéo produit une forme abrégée de la vidéo originale en extrayant le contenu le plus important et pertinent. Le résumé vidéo peut être utilisé dans diverses applications, telles que les systèmes de recherche et de navigation interactive, les utilisateurs peuvent également l'utiliser pour gérer et accéder au contenu vidéo numérique [Lienhart et al., 1997] [Barbieri et al., 2003] [Money and Agius, 2007] [Lew et al., 2006].

L'aspect multimédia de la vidéo, comprenant des informations audio, des sons, de la musique, des informations visuelles, images fixes ou en mouvement, et parfois des informations textuelles [Dimitrova, 2004], induit une complexité supplémentaire pour le résumé vidéo par rapport au résumé textuel et à d'autres techniques de traitement de texte. En plus des fonctionnalités de bas niveau, le résumé vidéo doit tenir compte de la sémantique des informations présentes dans la vidéo pour faciliter la compréhension du résumé.

Alors que beaucoup d'efforts ont été consacrés à la construction du résumé d'une seule vidéo [Money and Agius, 2007] [Yahiaoui et al., 2001], moins d'attention a été accordée à la construction d'un résumé pour un ensemble de vidéos [Yahiaoui et al., 2001]. Avec l'augmentation de la quantité de vidéos, celles-ci sont très souvent organisées en groupes, par exemple, le site YouTube présente les vidéos connexes dans la même page Web. Par conséquent, la question de la création d'un résumé pour un ensemble de vidéos devient d'une grande importance, et Ceci suit une tendance qui est maintenant bien établie dans la communauté du language naturel. Le résumé vidéo multi-documents a été largement étudié [Wactlar, 2001] [Dumont and Merialdo, 2008] [Das and Martins, 2007] [Wang and Merialdo, 2009]. Cependant ils présentent de nombreuses limites. Beaucoup d'algorithmes existants ne considèrent que les caractéristiques de la partie vidéo, et négligent la piste audio en raison de la difficulté de fusioner les informations audio et vidéo. Plusieurs algorithmes récents [Sugano et al., 2002] [Furini and Ghini, 2006] [Xu et al., 2005] considèrent les pistes audio et vidéo, mais ils sont souvent spécifiques d'un domaine. Tels que la caractéristique de mouvement basé sur le MPEG-7 et la détection des instants marquants par l'analyse de classe audio et du niveau sonore qui sont les mesures les plus utilisées dans [Sugano et al., 2002]. Dans [Furini and Ghini, 2006] les auteurs ont supposé que les séquences vidéos qui correspondent aux silences ne sont pas pertinentes alors que dans certains cas ils peuvent bien l'être. Dans [Xu et al., 2005] l'auteur propose un algorithme pour le résumé de vidéo de scène musicale en détectant le chœur dans la composante audio et la répétition des prises dans le visuel. Ces trois approches proposées sont des exemples de l'utilisation combinée du contenu audio et vidéo mais ces approches demeurent fortement liees à leurs domaine d'applications. Jusqu'à présent il n'y a pas d'algorithme générique pour le résumé de vidéo ce qui s'explique par le fait qu'il est plus facile de se situer dans un domaine spécifique ou des informations a priori pourront être utilisées. Un exemple simple serait le cas des vidéos de sport où le cri des spectateurs est un indicateur sur impertance de la scène. Ce genre d'informations ne pouvant pas être utilisées dans le cas des algorithmes génériques explique clairement la difficulté de l'approche de résumé de vidéo d'une manière globale.

Dans cette thèse nous proposons un algorithme générique de résumé d'un ensemble de vidéoǎ: Video-MMR. L'idée de base de cette approche vient de l'approche MMR (Maximal Marginal Relevance) qui a été développée à la base pour le résumé de texte par [Carbonell and Goldstein, 1998]. Ensuite nous avons ajouté l'information audio pour développer les algorithmss Audio Vidéo MMR (AV-MMR), le Balanced AV-MMR et le Balanced AV-MMR optimisé étape par étape (Optimized Balanced AV-MMR). En ajoutant de la transcription de parole en texte nous avons proposé le Texte Video MMR (TV-MMR). En plus nous avons suggéré le choix dŠnombre optimal de trames dans l'espace stationnaire des trames et de texte par Video-MMR, et ensuite de la durée optimale pour les séquences par TV-MMR. Au-delà de ses améliorations de la méthode MMR en ajoutant du flux multimédia, nous avons amélioré le MMR en analysant ses caractéristiques et proposé le Video-MMR2 qui exploite l'information visuelle mais surpasse les qualités du Video-MMR.

Après avoir généré le résumé vidéo, nous avons besoin d'évaluer les performances de l'algorithme. Ce qui relève de l'évaluation des algorithmes de résumé de vidéo [Das and Martins, 2007] qui est un problème bien connu et demeure ouvert à de nouvelles idées. On peut facilement distinguer un bon résumé d'un mauvais résumé, mais le résumé idéal n'existe pas. Par conséquent il devient difficile de mesurer la qualité d'un résumé vidéo automatiquement par un processus autonome. Il est possible de faire une évaluation subjective du résumé vidéo, mais ce genre d'évaluation prend du temps, coûte cher et n'est pas facilement reproductible. Ceci a conduit au développement de plusieurs algorithmes basés sur les techniques d'apprentissage automatique. Une métrique d'évaluation autonome de la qualité de résumé qui serait fortement corrélée avec celle d'une évaluation humaine serait très intéressante. Des situations similaires ont été déjà rencontrées dans le domaine de de la traduction automatique [Nilsson et al., 2007] ou du résumé de texte [Lin and Hovy, 2003]. En s'inspirant des techniques développées dans ces domaines nous proposons dans cette thèse un algorithme performant d'évaluation de résumé de vidéo, le VERT (Video Evaluation by Relevant Threshold).

En général, la recherche dans cette thèse fait les contributions suivantes:

1. Algorithmes de résumé multi-vidéo par l'information visuelle: Vidéo-MMR et vidéo-MMR2.

2. Algorithmes de résumé multi-vidéo par l'information visuelle et acoustique: AV-MMR, Balanced AV-MMR et OB-MMR.

3. Algorithme de résumé multi-vidéo par l'information visuelle et textuelle: TV-MMR.

4. L'optimisation de la présentation résumé des résumés statiques et dynamiques.

5. Mesure d'évaluation automatique de résumés vidéo: VERT.

A.1 Vidéo-MMR

Différentes techniques de résumé de vidéo ont été proposées. Dans cette thèse nous proposons une nouvelle approche de résumé d'ensemble de vidéo très efficace. Le Vidéo Maximal Marginal Relevance (Video-MMR), qui est une extension d'un algorithme classique utilisé pour le résumé de texte, le Maximal Marginal Relevance (MMR). Video-MMR maximise les séquences vidéo pertinentes et réduit les séquences redondantes comme le MMR avec les fragments de texte.

A.1.1 MMR

L'Accès rapide à l'information s'accroit, les algorithmes traditionnels se focalisent plus sur la pertinence par rapport à la demande de l'utilisateur. Cependant, si l'information pertinente se trouve sur une très grande quantité de fichiers, ceci entraine une redondance de l'information. Ce qui implique la nécessité d'un algorithme de résumé de document ayant une propriété anti-redondance. Dans la communauté du traitement automatique du langage naturel (NLP), le résumé automatique de texte a été étudié et plusieurs algorithmes ont été proposés durant la moitié du siècle dernier. Le résumé de texte est défini comme "un texte qui a été produit à partir d'un ou de plusieurs textes, qui contient l'information la plus importante du ou des texte(s) original(aux) et qui fait moins de la moitié du ou des texte(s) original(aux) et généralement moins que cela" dans [Radev et al., 2002], et plusieurs termes ont été aussi définis comme: "extraction" définit comme le processus d'identification et de reproduction de la section contexte; la tâche d'"abstraction" pour générer un contenu significatif de manière differénte; "fusion" qui consiste à concaténer avec cohérence des extraits; La "compression" consiste à la suppression des parties non significatives du texte original. D'après ses différents termes, on peut avoir plusieurs type de résumé [Das and Martins, 2007]: "résumé par extraction" qui produit un résumé relativement aux phrases du texte; "résumé par abstraction" produit un résumé basé sur la grammaire à l'aide des techniques avancées de génération automatique de textes; cependant en recherche d'information (IR), "résumé en fonction du sujet" dépend de la préférence de l'utilisateur et du sujet qui l'intéresse. Pour le résumé d'un texte unique ces algorithmes utilisent les techniques bayésienne, arbre de décision, model de Markov caché, les modèls en log-linéaires, les réseaux de neurones et d'autres approches spécifiques par rapport au langage utilisé. Dans les approches courants [Das and Martins, 2007] pour le résumé multi-documents, il y a plusieurs approches efficaces teles que: l'abstraction et la fusion d'information, Maximum Marginal Relevance (MMR), graph spreading (étalement de graphe) activation, les approches basées sur les centroides et les approches multi-langages.

Jusqu'à présent l'approche la plus populaire et efficace est le MMR proposer par [Carbonell and Goldstein, 1998]. Les auteurs créent le résumé en subdivisant le document en passages (phrase dans notre cas) et puis réorganisent les passages par rapport à la demande de l'utilisateur ou d'un système par l'approche MMR avec la mesure de similarité en cosinus. Dans l'approche, le MMR favorise les parties qui sont pertinentes par rapport au sujet de la même façon que pour les documents présentant des informations nouvelles non encore sélectionnées.

Dans [Carbonell and Goldstein, 1998], les auteurs indiquent que le MMR est plus efficace sur les documents assez longs et très utile pour l'extraction de passages sur un sujet dans plusieurs documents. Comme les nouveaux documents contiennent beaucoup de répétitions, les auteurs ont montré que les 10 premiers passages contiennent beaucoup de répétitions avec les méthodes précédentes, cependant le MMR peut réduire voire supprimer toutes les redondances.

A.1.2 Vidéo-MMR

Le but du résumé vidéo est d'identifier des trames ou séquences vidéos qui contiennent la majorité de l'information contenue dans la vidéo. Une séquence vidéo peut correspondre à une ou plusieurs trames donc ici nous nous sommes basés sur la sélection des trames. Comme le résumé vidéo a des similarités avec le résumé de texte nous proposons d'adapter les critères du MMR aux caractéristiques du vidéo pour construire un nouvee algorithme de résumé de Vidéo, le Vidéo-MMR [Li and Merialdo, 2010e].

En sélectionnant de manière itérative les trames pour construire le résumé, nous voudrions choisir celles dont le contenu visuel est similair du contenu des vidéos mais qui soit aussi différents au contenu des trames déjà sélectionnées dans le résumé. Cette approche est illustrée sur la Fig. A.1. En se basant sur cette mesure, un résumé S_{k+1} peut être construit itérativement en sélectionnant les trames à l'aide du Vidéo-MMR:

$$S_{k+1} = S_k \bigcup \underset{f_i \in V \setminus S_k}{\arg\max} \{\lambda Sim_1(f_i, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} Sim_2(f_i, g)\} \tag{A.1}$$

En se basant sur la définition du Vidéo-MMR, la procédure de résumé est décrite comme suit :

1. Le résumé vidéo initial S_1 est initialisé avec une image f_1, définie comme:

$$f_1 = \underset{f_i, f_i \neq f_j}{\arg\max} \prod_{j=1}^{n} Sim(f_i, f_j)^{1/n} \tag{A.2}$$

   Où f_j et f_i sont des trames de l'ensemble V de toutes les trames de toutes les vidéos et N est le nombre total de trame à l'exception de f_i.

2. Sélection de la trame f_k par Vidéo-MMR:

$$f_k = \underset{f_i \in V \setminus S_{k-1}}{\arg\max} (\lambda Sim_1(f_i, V \setminus S_{k-1}) - (1-\lambda) \max_{g \in S_{k-1}} Sim_2(f_i, g)) \tag{A.3}$$

3. L'ensemble $S_k = S_{k-1} \bigcup \{f_k\}$.

4. Itération à partir de l'étape 2 jusqu'à ce que S_k est atteint la taille désirée.

A.2 Vidéo-MMR2

L'approche la plus efficace exploitant seulement l'information visuelle est toujours importante pour le résumé vidéo, parce que l'information visuelle est l'information la plus significative et claire dans trois types d'information, textuelle, visuelle et audio. Donc, nous détaillerons plus notre algorithme, Vidéo-MMR, et essayons de l'améliorer après la discussion de ses limites.

Durant les expériences de Vidéo-MMR, nous trouvons qu'elle ne peut pas travailler efficacement quand plusieurs situations particulières se produisent :

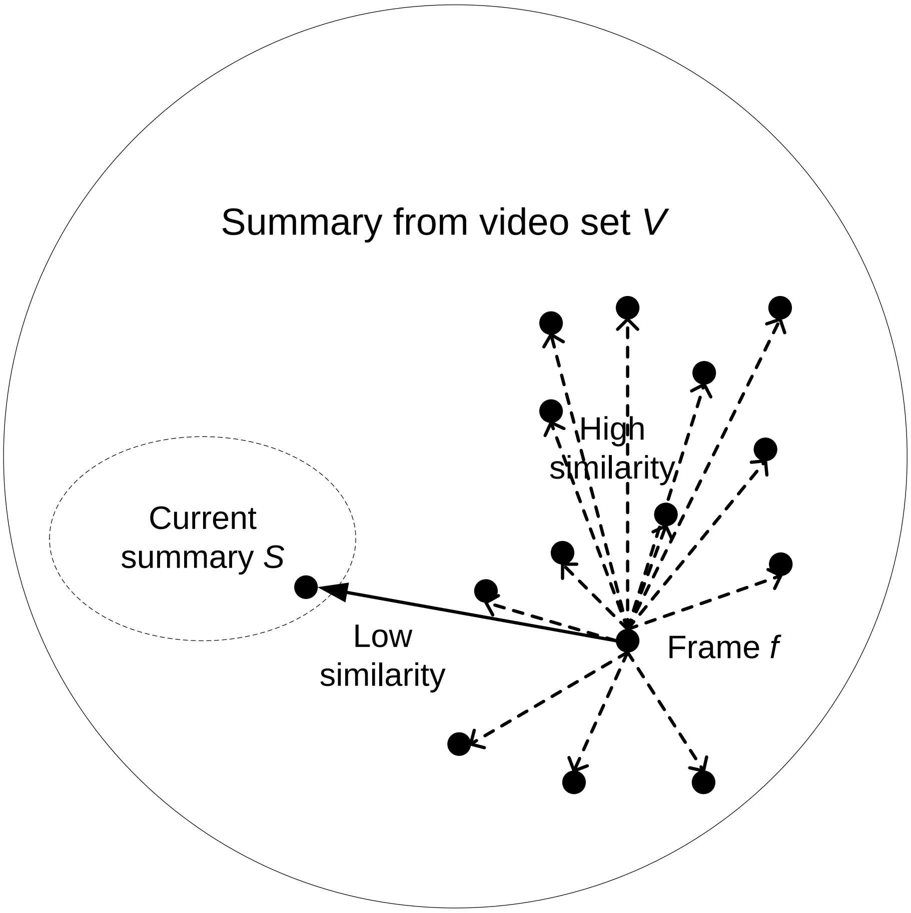
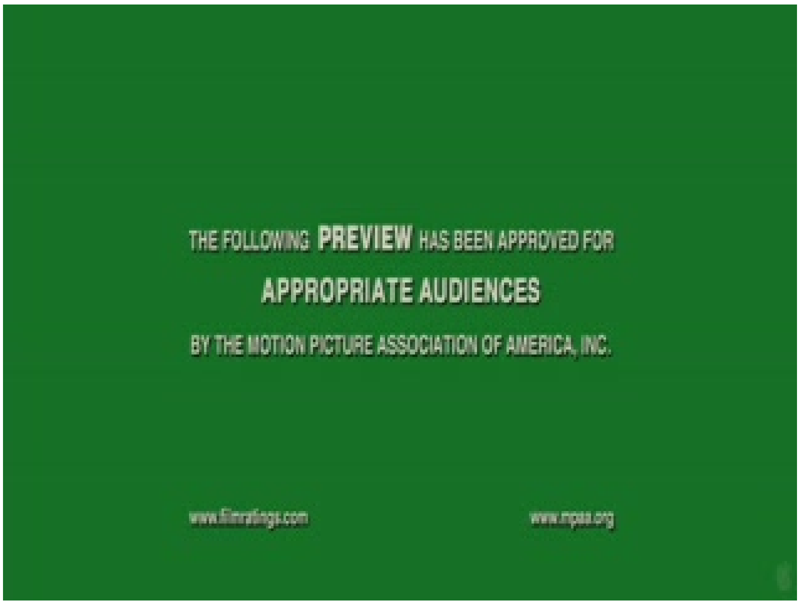


Figure A.1: L'illustration de la Vidéo-MMR



THE FOLLOWING PREVIEW HAS BEEN APPROVED FOR
APPROPRIATE AUDIENCES
BY THE MOTION PICTURE ASSOCIATION OF AMERICA, INC.

www.filmratings.com www.mpaa.org

Figure A.2: Les frame avec une couleur dominante dans la couleur verte

1. **La frame avec couleur dominante** Lorsqu'une grande partie des frames sont de la même couleur, comme la couleur verte dans les frames de début et de fin de plusieurs bandes annonces et la couleur sombre dans le début des frames d'une vidéo du journal. Ils sont présentés dans les figures Fig. A.2 et Fig. A.3.

   Les frames de Fig. A.2 et Fig. A.3 peuvent être choisies durant la procédure de la vidéo MMR comme il est montré dans la Fig. A.4, mais elles sont visuellement négligeables et dé inntile ées par l'utilisateur, bien que le poids de leurs mots de texte joue un rôle important dans un autre sens.

2. **Ls frame terne** Dans certaines frames, le contenu du frame est intéressant pour l'utilisateur, mais il est très terne pour distinguer ses détails. Un exemple de frame terne est présenté dans la figure Fig. A.5, qui est une frame dans le résumé montre dans la figure Fig. A.6. Il est difficile pour l'utilisateur d'examiner le contenu représenté dans ce genre d'images. Par conséquent, il vaut mieux éviter de sélectionner ces frames dans le résumé.

3. **La frame dupliquée** La dernière et la plus importante limite de Vidéo-MMR est le fait qu'elle peut sélectionner des frames dupliquées quand il y a beaucoup de frames dans la vidéo originale. La valeur de Sim_1 dans la formule de la Vidéo-MMR pour une frame augmente avec l'augmentation du nombre des frames dupliquées ou pour une frame très semblable pour lui.

   A partir de la figure Fig A.6, nous pouvons constater qu'il y a des frames similaires avec le visage de la même personne dans le résumé. Et nous affichons un exemple extrême dans la figure Fig. A.7 qui est créé par la Vidéo-MMR à partir des frames originaux présentés dans la figure Fig. A.8.

   Pour le résumé montré dans la figure Fig A.7, il est raisonnable d'inclure certains doublons pour un résumé de la taille de 10, parce que les frames individueles dans la figure Fig. A.8 sont 6. Mais les problèmes qui se posent sont :

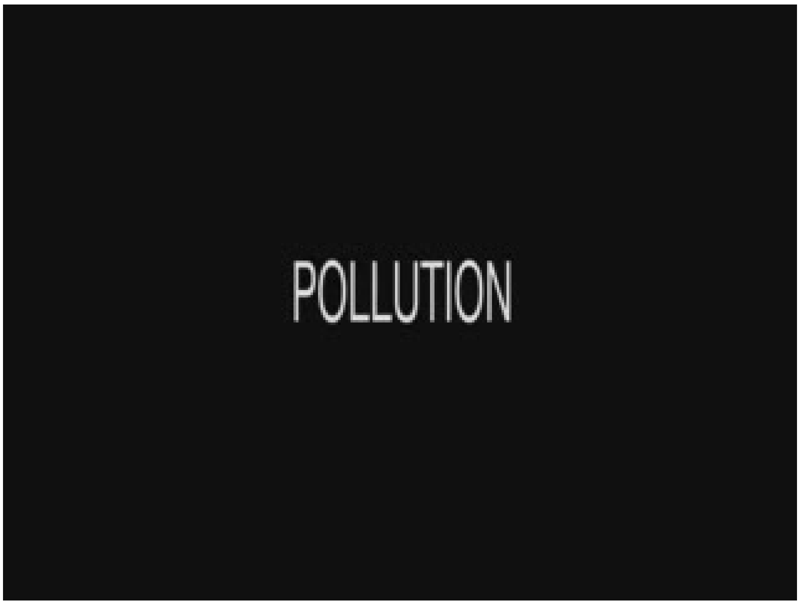


Figure A.3: Les frame avec une couleur majeure dans la couleur noire

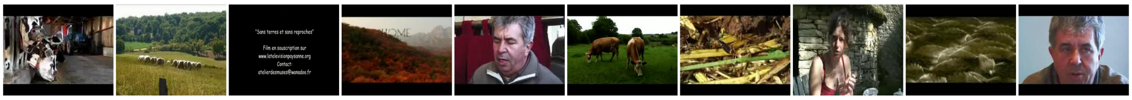


Figure A.4: Un résumé avec un frame avec une couleur majeure en noir

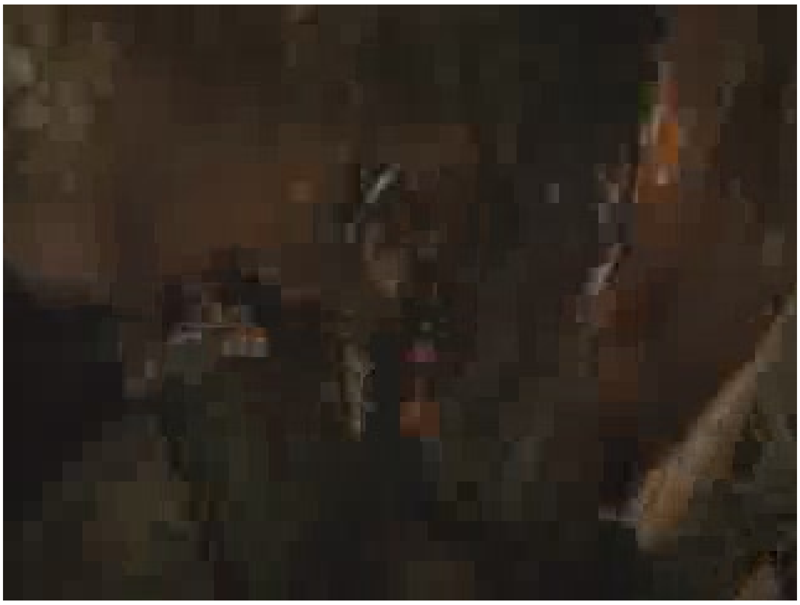


Figure A.5: Un frame ternes

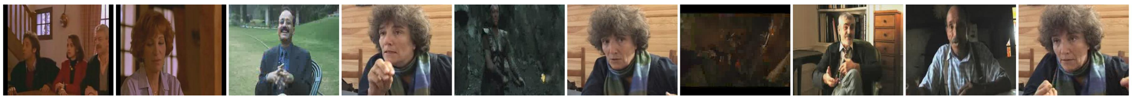


Figure A.6: Un résumé avec un frame ternes

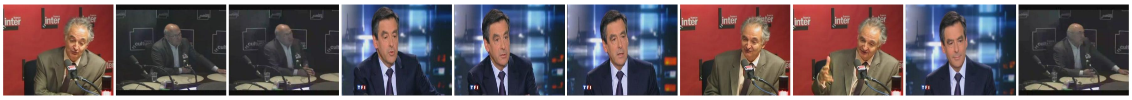


Figure A.7: Un résumé avec les doublons

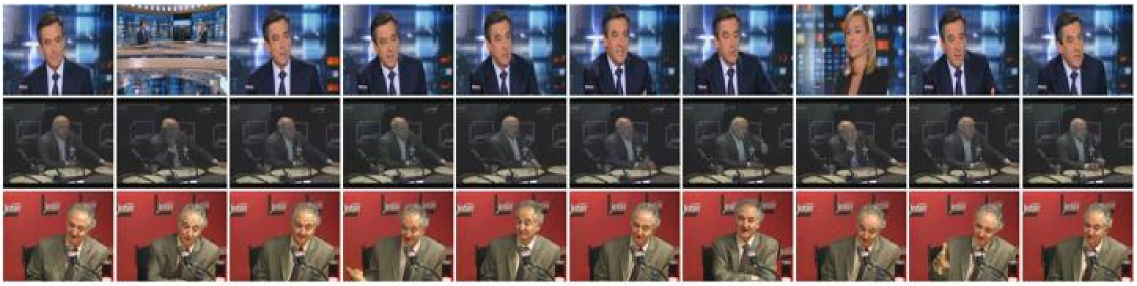


Figure A.8: Les frames d'origine pour Fig. A.7

- Les dupliqués sont choisis en premier, ce ne sont pas les frames nouvelles pour le résumé dans le mécanisme incrémental de la Vidéo-MMR.
- Il provoque aussi que certains frames nouvelles pour le résumé de la figure Fig. A.8 ne sont pas choisies dans le résumé.

Totalement, nous avons besoin d'améliorer la Vidéo-MMR dans ces trois aspects, en particulier les dupliqués. Nous proposons Vidéo-MMR comme suit :

- On évite les frames avec un couleur majeure, pour les éviter, nous utilisons tout simplement l'histogramme de couleur.
- Eviter les frames ternes. Pour les supprimer de frames candidats pour le résumé, nous décidons d'exclure tout simplement les frames ayant une valeur extrême faible pour le canal des valeurs dans l'histogramme de HSV. En imitant la méthode d'exclure les frames avec une couleur majeure, nous calculons d'abord l'histogramme de HSV d'un frame de 16 bins pour chaque canal. Ensuite, les frames avec des petites valeurs sont exclues de l'ensemble des trames candidats.
- Supprimer les dupliqués. Le facteur le plus important qui diminue la qualité du résumé est la frame dupliquée. L'approche pour éviter du duplicata est de réduire, Sim_1 en Vidéo-MMR, l'importance du frame avec une large quantité des trames identiques ou similaires. Par conséquent, nous proposons Vidéo-MMR2.

La procédure de la vidéo MMR2 résumé est décrite dans les étapes suivantes:

1. Exclure les frames visuellement sans importance de V.
2. le premier vidéo résumé S_1 est initialisé avec un seul frame f_1, défini comme suit :

$$f_1 = \arg\max_{f_i, f_i \neq f_j} (\prod_{j=1}^{n} Sim(f_i, f_j))^{1/n} \tag{A.4}$$

   où f_j et f_i sont des frames de l'ensemble V de toutes les trames de toutes les vidéos, et N est le nombre total des frames sauf f_i.

3. Sélectionner la frame f_k par la vidéo-MMR2.

4. L'ensemble $S_k = S_{k-1} \bigcup \{f_k\}$.

5. Itérer à l'étape 3 jusqu'à ce que S_k atteint la taille désirée

A.3 Multimédia MMR

Dans la section précédente nous avons présenté le principe du Video-MMR. Cependant le Vidéo-MMR exploit seulement l'information visuelle de la vidéo, l'information audio et textuelle sont ignorées or elles demeurent significatives pour appréhender le contenu vidéo dans sa globalité. Nous avons donc proposé une extension du Vidéo-MMR en exploitant l'information audio et textuelle de la vidéo.

Dans cette section, nous ajoutant la composante audio dans la formule du Vidéo-MMR en imitant l'approche Vidéo-MMR et proposons l'AV-MMR (Audio Vidéo MMR). Cependant une simple extension du Vidéo-MMR avec les caractéristiques de l'audio ignore la relation existant au niveau de la sémantique entre la composante audio et vidéo. En tenant en compte de la relation entre les composantes audio et vidéo et l'information temporelle de l'ensemble des vidéos nous proposons le Balanced AV-MMR. Dans le Balanced AV-MMR plusieurs paramètres sont manuellement choisis et donc nous proposons un algorithme avec une optimisation automatique des paramètres dans le OB-MMR (Optimized Balanced Audio Vidéo MMR).

A.3.1 AV-MMR

La plupart des algorithmes de résumé de vidéos exploitent l'information de la composante visuelle dans la vidéo car il est difficile de combiner différent type de flux multimédia. Cependant ceci limite pour plusieurs vidéos l'exploitation de l'information textuelle ou audio. Nous avons fait une extension du Vidéo-MMR pour tenir compte de l'information audio et visuelle dans l'AV-MMR. Pour chaque trame vidéo nous associons un deuxième segment audio correspondant à la trame. Ainsi nous exploitons les caractéristiques audio et vidéo et introduit une extension du Vidéo-MMR. Le résumé par l'approche AV-MMR reste donc similaire à ceux du Vidéo-MMR.

A.3.2 Balanced AV-MMR

AV-MMR est une extension directe de Vidéo-MMR et ne prend pas en compte les caractéristiques spécifiques de la bande audio et vidéo. En conséquence, nous proposons dans cette section un algorithme plus évolué, appelé Balanced AV-MMR ou BAV-MMR (Balanced Audio Vidéo Maximal Marginal Relevance) [Li and Merialdo, 2012], pour améliorer AV-MMR par:

1. La prise en compte de l'équilibre entre les informations auditives et les informations visuelles dans de courtes périodes;

2. L'analyse et l'utilisation des genres audio;

3. L'exploitation des changements d'un genre audio à un autre;

4. L'analyse et l'utilisation de l'information apportée par le visage;

5. L'utilisation de la "distance temporelle" dans un ensemble de séquences vidéo;
6. Enfin le design d'un nouveau mécanisme pour combiner ces propriétés.

Supposons que dans une courte période, le son attire plus l'attention de l'utilisateur, celui-ci ferait moins attention au contenu vidéo, et vice versa, car l'attention d'une personne pendant une courte période est limitée. Dans un segment audio, la durée est habituellement courte. Ainsi, il y a un équilibre entre l'information audio et l'information vidéo dans un segment audio. En conséquent, nous donnons à notre nouvel algorithme le qualificatif "balance". Balanced AV-MMR exploite l'information de différent genre audio, du visage et du temps pour augmenter l'équilibre de l'information et les similarités des séquences au niveau sémantique. Le genre audio et le visage sont des traits important d'une vidéo qui peuvent influencer l'équilibre entre l'audio et la vidéo dans un segment audio. Quand une transition audio se produit, il y a un changement audio significatif. A ce moment, un utilisateur fera plus attention à l'audio et l'audio devient plus important que normalement dans l'équilibre. Similairement, quand un visage apparait dans la bande vidéo d'un segment audio, le contenu vidéo devient plus important dans l'équilibre. De plus, le visage et le genre audio peuvent influencer les similarités entre les séquences au niveau sémantique. Pour une séquence vidéo, la similarité entre deux séquences contenant toutes deux un visage est plus grande que la similarité entre une séquence avec un visage et l'autre sans. Pour la bande sonore, deux séquences du même genre audio, par exemple la parole, sont plus similaires. Dans une vidéo, deux images proches dans le temps apparaissent redondantes. Deux images dans une vidéo ont l'air moins différent que deux images provenant de deux vidéos individuelles et non-dupliquées, même si celles-ci ont les mêmes similarités vis-à-vis des caractéristiques bas niveaux. Ainsi, il est nécessaire de considérer l'influence de l'information temporelle dans notre résumé.

1. Fundamental Balanced AV-MMR. A partir de la formule de AV-MMR et de l'analyse de l'équilibre entre l'information audio et vidéo dans un segment, nous introduisons un facteur d'équilibre entre l'information visuelle et auditive.
2. Balanced AV-MMR V1: utilisant le genre audio. Par l'analyse audio, nous savons que le genre audio est un trait important et peut refléter les caractéristiques de la vidéo. Il est clair que des extraits sonores ("audio frames") d'un même genre sont plus similaires que des extraits sonores de genre différents, même si elles ont la même similarité selon des critères audio comme les "Mel-frequency cepstral coefficients" (MFCC). MFCC est utilisé pour obtenir la similarité de la puissance court-terme de deux extraits sonores comme AV-MMR, mais la similarité au niveau sémantique ne peut pas être reflétée. En conséquence, nous pouvons introduire un "augment" facteur pour les genres audio pour ajuster la similarité des vecteurs MFCC. Les transitions audio indiquent des changements audio significatifs. Dans la catégorie *Musique*, la transition du silence ou de la musique à la parole indique l'apparition possible d'un chanteur qui commence à chanter à ce moment. Dans la catégorie *Nouvelles*, la transition du silence à la parole indiques habituellement le début des nouvelles par le journaliste ou le présentateur. Autours de du changement audio, l'utilisateur fera plus attention à l'audio et moins attention à la bande vidéo selon notre principe d'équilibre.
3. Balanced AV-MMR V2: utilisant la détection de visage. Le visage est extrêmement important dans l'information visuelle, donc l'image video devient plus importante

quand un visage apparait dans celle-ci. Comme notre principe d'équilibre favorise un parti et défavorise l'autre, le facteur d'équilibre devrait augmenter dans ce cas-là. En plus de notre facteur de balance, l'apparition du visage augment aussi la similarité entre 2 images video. Au niveau sémantique, une image contenant un visage est plus similaire qu'une autre image avec le visage qu'une autre image sans celui-ci. De plus, deux images avec des visages reflètent souvent le contenu pertinent d'une vidéo, comme plusieurs journalistes dans *Nouvelles* et acteurs dans *Film*.

4. Balanced AV-MMR V3: ajout du facteur de distance temporelle. Enfin, nous considérons l'influence de la distance temporelle de deux images f_i et f_j, d'une même vidéo ou non, sur la similarité visuelle et auditive :

   - Des images d'une video plus proches dans le temps ont habituellement un contenu plus pertinent et donc deux images plus proches dans une video sont reconnues comme plus similaires.
   - Pour des videos multiples, une image est plus similaire d'une autre image dans la même video qu'une image provenant d'une autre video non-dupliquée.

   Il est ensuite possible de prendre en compte dans le résumé l'information temporelle pour sélectionner des images de multiples videos. Cet équilibre est appelé "équilibre temporel".

A.3.3 OB-MMR

Nous améliorons Balanced AV-MMR en optimisant les paramètres de l'algorithme, et proposons OB-MMR. Balanced AV-MMR exploite plusieurs paramètres qui sont réglés manuellement et empiriquement. Ces paramètres sont: le paramètre de balance entre l'information auditive et visuelle, le paramètre de distance temporelle, le paramètre de visage, le paramètre de genre auditif, et le paramètre de transition auditive. Cependant, il est difficile de décider manuellement des meilleures valeurs pour ces 5 paramètres. Aussi, pour différents genres de vidéos, les valeurs optimales de ces paramètres peut varier, parce que la relation entre la piste vidéo et audio est différente. Par conséquent, nous souhaitons proposer un mecanisme automatique pour optimiser l'ensemble de poids pour Balanced AV-MMR.

Il est alors nécessaire de sélectionner un algorithme automatique pour ajuster automatiquement les poids. Un algorithme efficace est l'Optimisation de Particule de Nuage (Particle Swarm Optimization, PSO) proposé par [Poli et al., 2007] [Kennedy and Eberhart, 1995]. PSO a été utilisé pour un large panel d'applications, qui ont prouvé l'effet de PSO. "Le nuage comme un tout, à la manière d'un troupeau d'oiseau provisionnant collectivement de la nourriture, est susceptible de se rapprocher d'un optimum de la fonction d'ajustement. Une particule en elle-même n'a pratiquement aucun pouvoir de résolution de problème; un progrès s'observe uniquement lorsque les particules interagissent." [Poli et al., 2007]. Dans PSO, chaque particule décide de son mouvement en considérant sa position courante et la précédente meilleure position des particules. Un individu contient trois vecteur D-dimensionnels: la position courante $\overrightarrow{x_i}$, la précédente meilleure position $\overrightarrow{p_i}$, et la vitesse $\overrightarrow{v_i}$. La meilleure fonction résultante est dénotée $pbest_i$, et $\overrightarrow{p_g}$ est le meilleur voisin de $\overrightarrow{p_i}$. La procédure PSO est décrite dans [Poli et al., 2007]. Dans OB-MMR, 5 poids peuvent être considérés comme 5 éléments d'un vecteur $\overrightarrow{x_i}$ dans l'espace de recherche de PSO pour la fonction d'ajustement.

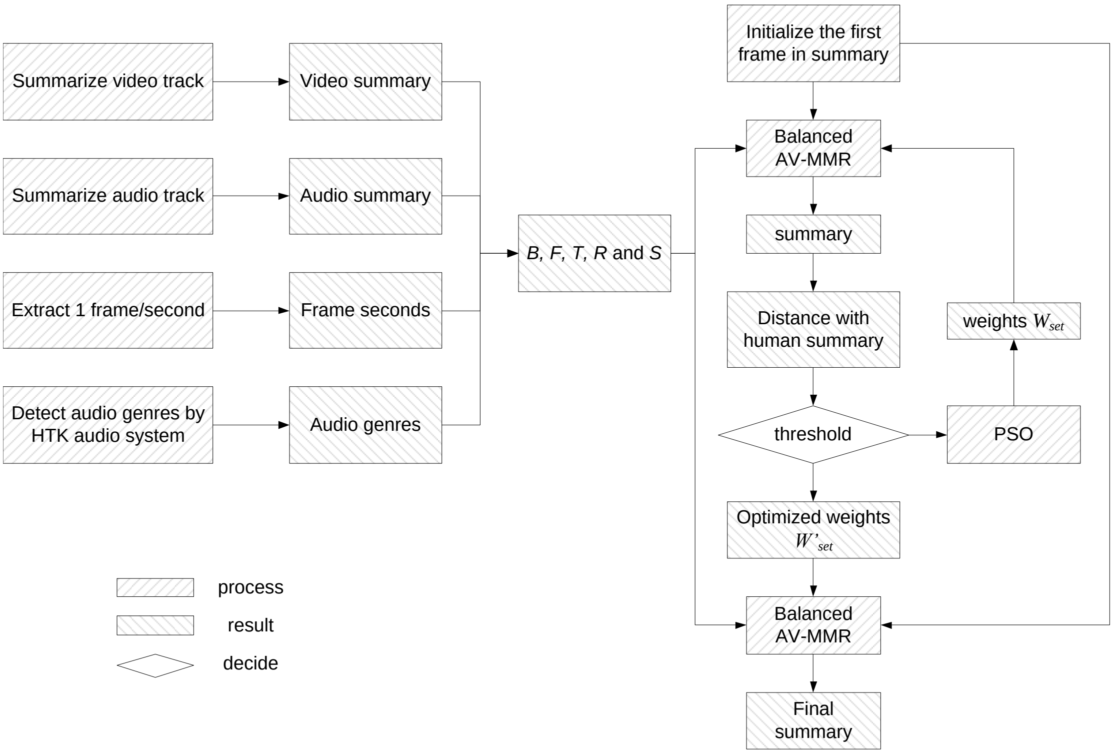


Figure A.9: Le cadre de OB-MMR

Un autre algorithme simple pour trouver les poids optimisés pour OB-MMR est Quadrillage et Détente (Gridding and Relaxation, GR). Ici, quadrillage signifie le quadrillage moyen des poids possibles sur une étendue adéquate, et l'essai de chaque combinaison des poids pour optimiser la fonction d'ajustement, la similarité entre le résumé OB-MMR et le résumé humain. Puisque l'intervallle entre deux valeurs de quadrillage est initialement élevée de sorte que le calcul soit suffisamment rapide, lorsque les meilleures valeurs sont trouvées sur la grille, un processus similaire est répété recursivement avec une grille plus fine autour du point optimal, ce qui est appelé l'étape de détente. La fonction d'ajustement pour quadrillage et détente est la même que précédemment, la similarité entre le résumé OB-MMR et le résumé humain. La détente se produit plusieurs fois, jusqu'à ce que l'intervalle de la grille atteigne la précision souhaitée.

Nous illustrons le cadre dans la figure Fig. A.9.

A.3.4 TV-MMR

Dans les sections précédentes, nous avons exploités deux informations du multimédia, la vidéo ainsi que l'audio afin d'améliorer la synthèse d'une vidéo. Cependant, les informations textuelles ont également une très grande importance dans les vidéos. Pour cette raison, nous proposons Text Vidéo MMR (TV-MMR) afin d'exploiter les informations textuelles et visuelles afin de synthétiser au mieux la vidéo. TV-MMR sélectionne les segments vidéo correspondant aux n-grammes en utilisant à la fois le texte et le contenu visuel. Dans Video-MMR, l'unité d'information de base est une image-clé unique, alors que dans TV-MMR on parle de segment n-grammes. Le contenu visuel d'un segment n-grammes est composé de plusieurs images-clé qui apparaissent entre le début et la fin de l'énoncé. Afin d'accélérer le calcul, nous proposons une vidéo synthèse de une image par seconde, ainsi une vidéo de 5 secondes sera représentée par 5 images-clés. La similarité entre les images-clés utilisée dans Video-MMR est étendue à une similarité entre un ensemble d'images clés en calculant la moyenne des similarités des images-clés.

A.3.5 Résumés statique et dynamique

On trouve deux types de résumé vidéo: résumé stationnaire et résumé dynamique. Dans cette section, nous proposons deux approches d'optimisation pour ces deux types de résumés en basant sur la vidéo et le texte. La première approche met la lumière sur les résumés statiques (résumé stationnaire), où le résumé est considéré comme un ensemble d'images et de mots clés sélectionnés, afin qu'ils soient affichés dans une zone fixe. Alors que la seconde approche traite les résumés dynamiques (écrémé vidéo) où les segments d'une vidéo sont sélectionnés en basant à la fois sur leur information visuelle et le contenu textuel de l'ASR pour composer une nouvelle séquence vidéo de durée prédéfinie. L'approche pour le résumé statique s'appuie sur la Vidéo-MMR, alors que l'approche pour un résumé dynamique a besoin d'utiliser la TV-MMR.

Un résumé statique est essentiellement composé d'un ensemble d'images-clés sélectionnés. Néanmoins, il est important d'utiliser également une partie de l'espace d'affichage afin de montrer certains mots-clés qui sont liés au contenu de la séquence vidéo. Dans notre travail, nous avons recours à la transcription de la parole de la piste audio comme il est décrit. Le résumé est souvent présenté dans un espace d'affichage avec une taille prédéfinie, tel est l'exemple d'une page web. C'est pour cela que l'algorithme du résumé doit sélectionner un nombre prédéfini d'images-clés pour qu'elles s'adaptent à cette espace, tout en maximisant la quantité d'informations présentant à l'utilisateur. Quand des mots-clés ou des phrases sont aussi disponibles, l'algorithme du résumé doit décider, non seulement à propos des mots clés à afficher, mais aussi sur le nombre relatif de mots-clés et des images-clés pour tenir dans l'espace prédéfini.

Notre résumé dynamique est une concaténation des segments audio visuels extraits à partir des vidéos originaux. Les segments candidats qui ne sont pas sélectionnés sont ceux correspondant aux énoncés de n-grammes. La durée d'un segment coïncide avec l'énoncé de n-grammes du texte. Dans cette section, nous discutons seulement les résumés dynamiques du point de vue de maximiser l'information dans les résumés, bien que le flot de l'histoire est aussi important pour le résumé dynamique.

Une difficulté particulière vient du fait que le taux de flux d'information est différent entre les données audio et le canal visuel. Pour la partie visuelle, les vidéos sont une succession de coups. Ces derniers sont souvent assez longs (de l'ordre de 10 secondes ou plus), avec un mouvement lent (à l'exception des clips musiques). Dans ce cas, une présentation visuelle de 1 ou 2 secondes du coup est suffisante pour transmettre la plupart du contenu visuel du coup. Toute présentation plus longue est un usage inutile de la chaîne d'information visuelle pour le résumé. Par contre, pour la partie textuelle, la redondance est extrêmement rare, de telle sorte que les extraits plus longs fournissent plus d'information de contenu.

Par conséquent, le choix de la durée optimale de n-grammes est dirigé par deux contraintes opposées :

- Les valeurs plus petites de n favorisent plus le contenu visuel pour être présenté (pour une durée de résumé donnée);

- Les valeurs plus élevées de n permettent plus de cohérence dans l'information textuelle pour être inclus.

A.4 VERT

Le résumé vidéo est devenu un outil important pour le traitement de l'information multimédia, mais l'évaluation automatique d'un système de résumé vidéo reste un défi. Les gens peuvent être facilement distinguer entre "bon" et "mauvais" résumés, mais le meilleur résumé idéal n'existe pas, c'est pour cela qu'il est difficile de de définir une mesure de qualité qui peut être calculée automatiquement. Il est encore possible de mettre en place des expériences impliquant les hommes pour évaluer des résumés vidéos, mais ces expériences sont coûteuses, chronophage, et ne peuvent pas être facilement répétées, ce qui altère le développement de plusieurs algorithmes basés sur les techniques de la machine d'apprentissage. Une bonne mesure de qualité capable de réaliser un calcul automatique et de montrer une forte corrélation avec l'évaluation humaine est donc un grand intérêt.

Une situation similaire augmente dans le domaine de la machine traduction et résumé texte, où des spécifiques procédures automatiques, respectivement BLEU et ROUGE, évaluent la qualité d'un candidat en comparant ses similaires locaux avec plusieurs références générées par les hommes. Ces procédures sont maintenant couramment utilisées dans différentes benchmarks. Dans ce chapitre, nous développons cette idée dans le domaine de la vidéo et nous proposons l'algorithme VERT [Li and Merialdo, 2010c] [Li and Merialdo, 2010d] pour évaluer automatiquement la qualité des résumés vidéo. VERT, adapté à la fois pour un seul ou plusieurs vidéos, imite les théories du BLEU et ROUGE, et compte le nombre pondéré des unités sélectionnées qui se chevauchent entre le résumé vidéo généré par ordinateur et plusieurs artificielles références. Plusieurs variantes du VERT sont proposées et calculées. Rouge, vert et bleu sont les trois couleurs primaires peuvent devenir un ensemble d'algorithmes d'évaluation de référence dans leurs propres domaines.

A.4.1 BLEU

L'évaluation humaine de traduction automatique (MT) doit prend en compte de nombreux facteurs: adéquation, fidélité et aisance de la traduction [Hovy, 1999]. Les évaluations humaines sont parfaites mais prennent beaucoup de temps, parfois plusieurs semaines voire plusieurs mois. C'est pourquoi une méthode automatique d'évaluation de MT peut aider les chercheurs à évaluer rapidement leurs techniques de MT. Pour évaluer automatiquement la qualité d'une traduction automatique le BiLingual Evaluation Understudy (BLEU) [Papineni et al., 2002], basé sur la mesure de score de cooccurrence de n-grammes, a été proposé. BLEU était la métrique utilisée par le NIST (NIST 2002) pour classer les traductions. L'idée principale derrière BLEU est de mesurer les similarités entre une traduction candidate et un ensemble de traductions références.

Puisqu'une traduction par un professionnel est considérée comme idéale les traductions automatiques devraient s'en approcher le plus possible. Ainsi une MT sera meilleure si elle partage plus de mots et de phrases avec les traductions de références. C'est pourquoi BLEU devrait être capable de compter les n-grammes qui sont à la fois dans la traduction candidate et dans les traductions humaines de références. BLEU est une méthode de précision qui compte le nombre d'occurrences de n-grammes dans les traductions références et divise le nombre total de n-grammes dans la traduction à évaluer. Ce calcul est fait phrase par phrase. Les résultats obtenus par BLEU se sont avérés être très proche des évaluations faites par l'homme.

A.4.2 ROUGE

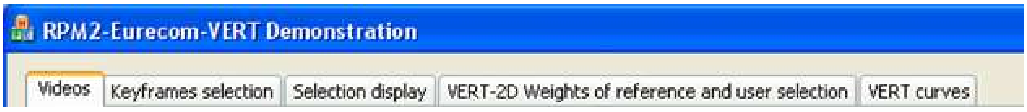
En ce qui concerne l'évaluation de résumé de textes, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) propose par Lin [Lin and Hovy, 2003] [Lin, 2004] détermine automatiquement la qualité d'un résumé par rapport à des résumés fait par l'homme. ROUGE compte le nombre d'occurrences de n-grammes, séquences de mots et paires de mots qui apparaissent dans les deux résumés. ROUGE a été utilisé dans la Document Understanding Conference (DUC), une évaluation à grande échelle de méthode de résumé, financé par NIST. Dans [Lin, 2004], plusieurs variantes sont proposées: ROUGE-N, ROUGE-L, ROUGE-W et ROUGE-S.

- ROUGE-N compte les n-grammes communs au résumé à évaluer et aux résumés de références.
- ROUGE-L donne plus d'importance à la plus longue séquence commune (LCS) entre le résumé à évaluer et les référénces. Plus la LCS de deux résumés est longue plus les résumés sont proches.
- ROUGE-W privilégie la plus longue séquence commune pondérée (WLCS), il diffère de ROUGE-L qui ne prend pas en compte les relations spatiales entre différentes LCS. Les LCS pondérées (WLCS) sont une amélioration des LCS car elles rendent compte de la longueur des cooccurrences consécutives des deux résumés.
- ROUGE-S est le Skip-Bigram Co-occurrence Statistics, un "skip-bigram" est une paire de mots quelconque mais dans le même ordre, ce qui permet de ne pas prendre en compte des parties de phrase de longueur arbitraire.

A.4.3 VERT

En reprenant les idées de ROUGE et BLEU, on peut adapter ces techniques de mesure aux résumés vidéos. Notre approche est de se concentrer sur la sélection d'images-clé, puisqu'une vidéo peut facilement être créée par la concaténation d'extraits vidéos autour des images-clé sélectionnées. Puisque nous pensons que la chronologie des images-clé n'est pas aussi importante que l'ordre des mots dans une phrase nous préférons classer en fonction de l'importance des images-clé dans la sélection. La création du résumé vidéo se fait donc ainsi:

1. On considère un ensemble de séquences vidéos V_1, V_2, ..., V_k correspondant à un certain sujet.
2. Ces séquences sont segmentées en plusieurs parties ou sous parties, chaque partie étant représentée par une ou plusieurs images-clé.
3. A l'aide des parties, sous parties ou images-clé on créée une sélection à inclure dans le résumé. Enfin, cette sélection peut être réordonnée pour mettre le contenu le plus important au début.
4. Le contenu sélectionné est assemblé dans un résumé vidéo, soit sous la forme d'un album images-clé ou une écrémé vidéo.



RPM2-Eurecom-VERT Demonstration
Videos
Keyframes selection
Selection display
VERT-2D Weights of reference and user selection
VERT curves

Figure A.10: Onglets du démonstrateur VERT

Une fois la sélection faite une pondération $w_S(f)$ liée au classement de l'importance de l'image-clé est attribuée à chaque image-clé f de la sélection S. Ainsi notre mesure VERT compare un ensemble d'images-clé sélectionnées par ordinateur avec un ensemble d'images-clés référence, sélectionné par l'homme. Puisque BLEU est une mesure de précision et ROUGE une mesure de cooccurrence nous proposons également VERT-Precision (VERT-P) et VERT-Recall (VERT-R) respectivement. Nous allons prouver que VERT-R donne de meilleurs résultats.

A.4.4 Le démonstrateur de VERT

Dans cette section nous présentons un démonstrateur qui permet de visualiser le processus d'évaluation de VERT. Nous avons créé une interface pour que l'utilisateur puisse facilement regarder les vidéos et sélectionner le résumé candidat. Ensuite la sélection de l'utilisateur est comparée avec les références et le résultat de la comparaison est visualisé à l'aide d'une grille colorée, les pondérations VERT de la sélection de l'utilisateur et de la référence sont également représentées par deux grilles colorées. Finalement les résultats de VERT sont affichés sous forme de courbes.

Notre démonstrateur est constitué de boites de dialogues dans différents onglets. Il y a 5 onglets: le premier contient 6 vidéos, le deuxième permet de sélectionner les images-clé, le troisième affiche la sélection, l'onglet suivant donne la pondération "VERT-2D", et le dernier onglet contient les courbes VERT. Une illustration de ces onglets est donnée dans la Fig. A.10. Nous allons décrire ces onglets en détail dans la suite de cette section.

A.4.4.1 L'onglet "6 Vidéos"

Dans cet onglet de notre démonstrateur 6 vidéos de notre ensemble de vidéos préparées "YSL" sont lues.

1. La "vidéo 1" est une vidéo de 7 minutes et 38 secondes d'un défilé de mode féminine

2. La "vidéo 2" est une vidéo de 29 secondes. C'est une publicité pour du parfum.

3. La "vidéo 3" est une publicité pour une marque de saucisson. Elle dure 19 secondes.

4. La "vidéo 4" dure 2 minutes et 20 secondes, c'est un passage du journal télévisé de CNN à propos d'un créateur de mode français.

5. La "vidéo 5" est un passage de patinage artistique qui dure 3 minutes et 12 secondes.

6. La "vidéo 6" est une sorte d'interview d'un chanteur qui vient de Hong Kong et dure 3 minutes et 6 secondes.

Le format de cet onglet est présenté dans la figure suivante, Fig. A.11.

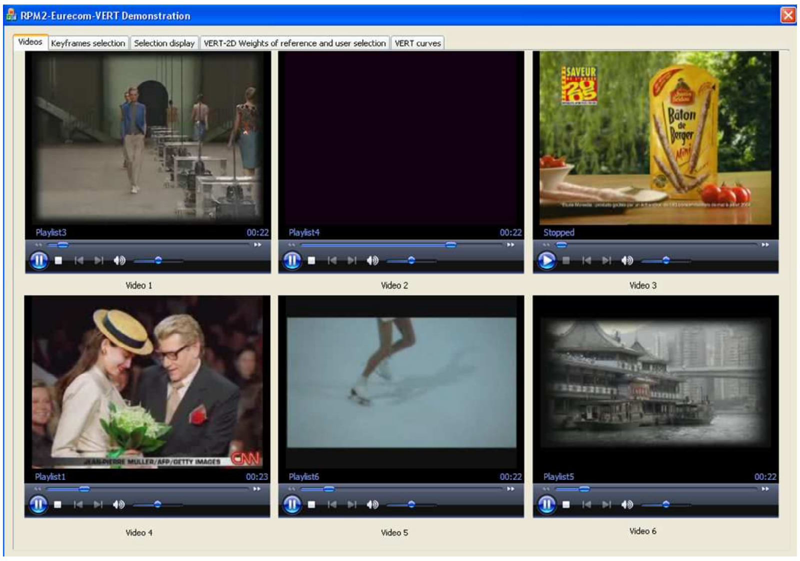


Figure A.11: L'onglet "6 Vidéos"

A.4.4.2 L'onglet "keyframes selection"

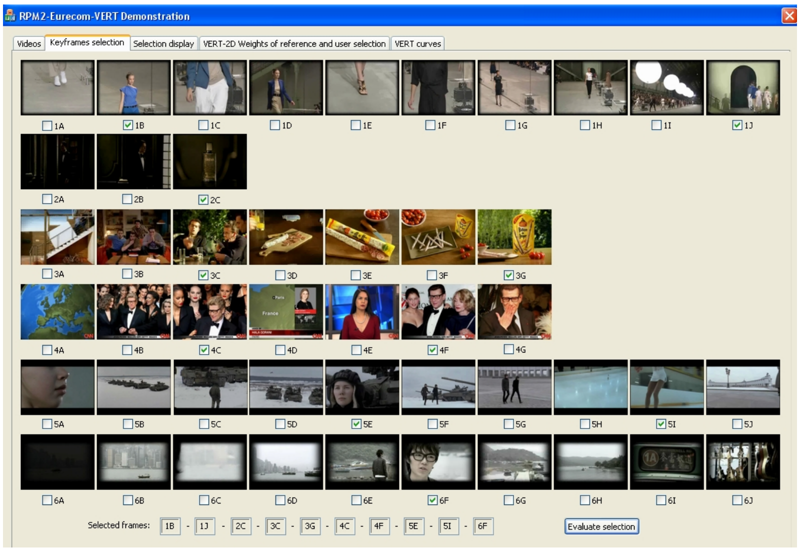
Cet onglet est composé de 47 images-clés présentées sur 6 lignes et 10 colonnes. Ces images-clés sont obtenues à l'aide de Video-MMR. Chaque ligne correspond à une vidéo et les colonnes représentent l'ordre chronologique. Dans la Fig. A.12, il y a une case à cocher sous chaque image-clé. L'utilisateur peut sélectionner 10 images clés puis cliquer pour terminer la sélection. Le choix du nombre de 10 images clés a été fait car nous avons un ensemble de référence d'images-clés, sélectionnés par 12 personnes et qui contiennent chacun 10 images clés. Après la sélection le démonstrateur va calculer les images de la démonstration et les courbes décrites dans les sous section A.4.4.3, A.4.4.4 et A.4.4.5.

A.4.4.3 L'onglet "selection show"

Quand l'utilisateur clique sur l'onglet "selection show", les images-clé qu'il a sélectionnées et les images clés de référence sont affichées dans une grille $13 * 47$. Le nombre 47 désigne le total des 47 images-clés dans l'ordre, de la première à la sixième vidéo. Ensuite de la première à la douzième ligne sont représentées les sélections des 12 personnes références. La dernière ligne de la grille est la sélection que l'utilisateur a créée dans l'onglet de sélection d'images-clé. L'ordre de sélection est représenté par des couleurs. Une barre de couleur est donnée sur la droite de la Fig. A.13. A partir de cela on peut lire les sélections de références et la sélection de l'utilisateur de manière intuitive.

A.4.4.4 L'onglet "VERT-2D weights of reference and user selection"

Cet onglet est compose de deux grilles 47*47. La taille 47 est choisie pour la même raison que précédemment. Puisque VERT-2D est un processus bidimensionnel et qu'il prend en compte l'ordre chronologique entre deux images-clé dans un 2-gramme les grilles ont deux dimensions. La grille de gauche représente les poids de références de VERT-2D dans VERT-R alors que la grille de droite représente les poids de la sélection de l'utilisateur de VERT-2D qui sont le numérateur. Le gradient de couleur de grande pondération à petite


RPM2-Eurecom-VERT Demonstration
Videos
Keyframes selection
Selection display
VERT-2D Weights of reference and user selection
VERT curves
Evaluate selection

□ 1A ☑ 1B □ 1C □ 1D □ 1E □ 1F □ 1G □ 1H □ 1I ☑ 1J

□ 2A □ 2B ☑ 2C

□ 3A □ 3B ☑ 3C □ 3D □ 3E □ 3F ☑ 3G

□ 4A □ 4B ☑ 4C □ 4D □ 4E ☑ 4F □ 4G

□ 5A □ 5B □ 5C □ 5D ☑ 5E □ 5F □ 5G □ 5H ☑ 5I □ 5J

□ 6A □ 6B □ 6C □ 6D □ 6E ☑ 6F □ 6G □ 6H □ 6I □ 6J

Selected frames: 1B - 1J - 2C - 3C - 3G - 4C - 4F - 5E - 5I - 6F

Figure A.12: L’onglet ”keyframes selection”

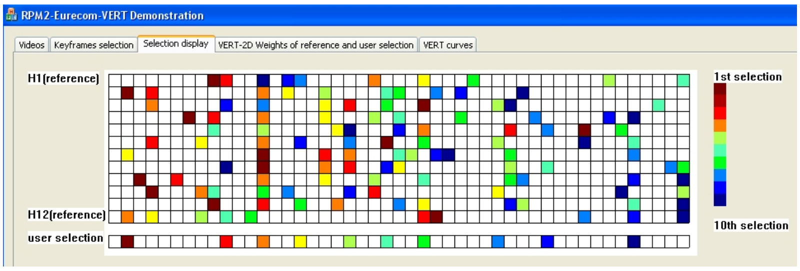


Figure A.13: L’onglet ”selection show”

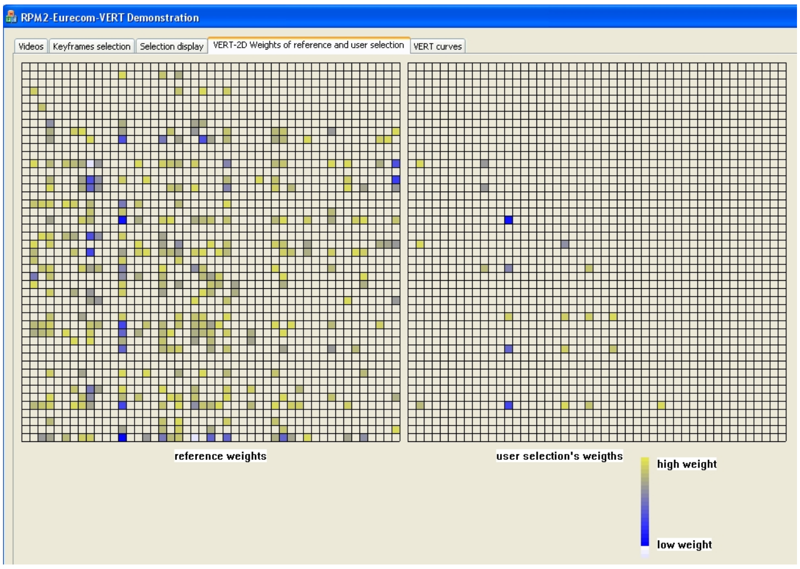


Figure A.14: L'onglet "VERT-2D weights of reference and user selection"

pondération est donné dans la Fig. A.14 également. Nous pouvons directement voir quels poids de VERT-2D sont sélectionnés par l'utilisateur.

A.4.4.5 L'onglet "VERT curves"

Fig. A.15 montre les courbes de VERT-2D, à l'aide de 12 points. Le point "U, H1" signifie que l'on utilize les 11 résumés références à part le premier comme nouvelles référence dans VERT-R.

A.5 Conclusion

La qualité de vidéos est devenue de plus en plus importante du jour à l'autre. Et le nombre des vidéos de l'internet, DV personnel, TV et des autres sources ne peut plus être géré par l'être humain. Par conséquent, le résumé vidéo est une façon très importante pour gérer les vidéos par sélection de l'information. Le résumé de vidéo est convenable pour les navigateurs et les moteurs de recherches, ça permet à l'utilisateur de faciliter la gestion et l'accès au contenu de la vidéo.

Dans cette thèse, on a proposé un résumé algorithme, Vidéo-MMR en utilisant l'information visuelle pour multi-vidéo. Vidéo-MMR est basé sur l'idée du résumé de texte. Il est aussi un algorithme incrémental pour reconstruire le résumé image par image.

On a développé Vidéo-MMR en adoptant deux façons:

- Une façon pour améliorer Vidéo-MMR est par examiner les avantages et éviter les défauts de Vidéo-MMR. Par conséquent, on a proposé Video-MMR2; qui a utilisé seulement l'information visuelle. Video-MMR2 évite de sélectionner le non-intéressant images le résumé de vidéo stationnaire, par incluant, l'image avec le couleur majoritaire and l'image terne. L'avantage de Video-MMR2 est d'éviter de sélectionner les images dupliquées dans le résumé.

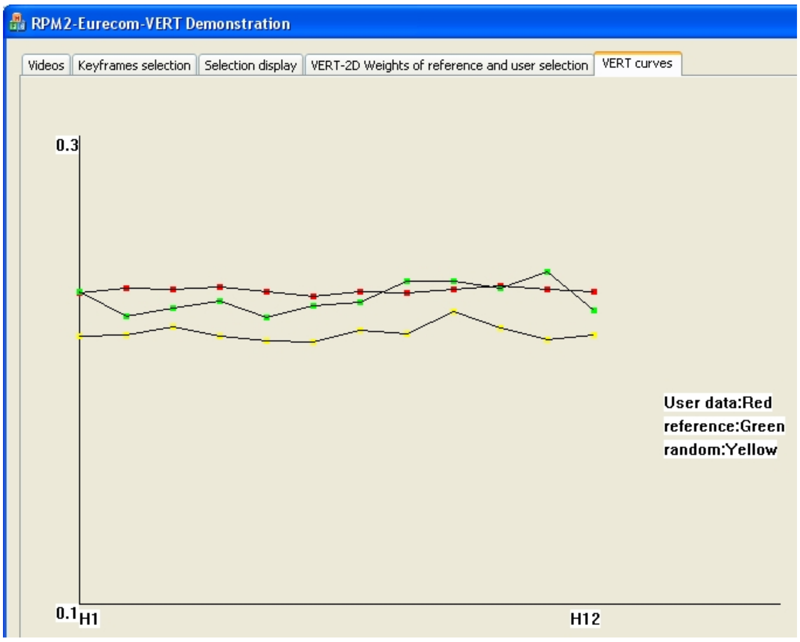


Figure A.15: L'onglet "VERT curves"

- On utilise l'information depuis l'audio ou le texte du vidéo. En introduisant l'information audio, on a proposé l'algorithme AV-MMR, Balanced AV-MMR et OB-MMR étape par étape. Sachant que l'information audio est très importante pour la vidéo, on a pu obtenir des meilleurs résultats que celles de Vidéo-MMR. Cependant, on a introduit le texte transcrit depuis la parole dans Vidéo-MMR pour former TV-MMR. En outre, pour les deux formes de résumé: statique et dynamique, on a proposé le nombre optimisé des images-clé avec un prédéfini stationnaire espace par Vidéo-MMR la durée du segment optimisé du résumé dynamique de la vidéo par TV-MMR.

L'algorithme du résumé vidéo est une seule facette dans le domaine, car l'évaluation de l'algorithme, notamment l'évaluation automatique est une autre intéressante facette. Avec une bonne évaluation, on ne peut pas décider la performance de notre nouvelle approche. Cependant, l'évaluation humaine est beaucoup mieux, mais elle prend beaucoup de temps, ce répète pas et elle est instable. Donc, une évaluation automatique est nécessaire pour le résumé des vidéos. Par conséquent, on a proposé VERT pour évaluer automatiquement the la qualité de résumé stationnaire. En utilisant VERT, on a besoin seulement de quelques résumés comme référence pour l'être humain; ensuite, on peut obtenir la qualité quantitative de diffèrent résumés générés par notre algorithme. VERT est un gain de temps, peut être répété and stable par rapport à l'évaluation humaine.

A.5.1 Perspective

Malgré les travaux déjà faites pour le résumé vidéo, il reste encore beaucoup de challenges et des améliorations dans des aspects différents [Money and Agius, 2007], qui peuvent être convenable pour notre travail aussi:

1. L'utilisation des traits sémantiques. Pour plusieurs algorithmes qui existent déjà dans le domaine de résumé vidéo, y compris l'algorithme proposé dans notre thèse,

il y a un manque d'utilisation de traits sémantiques. Donc ces résumes ne peuvent pas surmonter cette difficulté. En plus, la plupart des algorithmes ne peuvent pas générer un résumé personnalisé.

2. L'utilisation de traits externes et hybrides. En exploitant l'information fournie par l'utilisateur en dehors de la vidéo, ça peut faciliter la génération du résumé personnalisé. Par exemple, l'information de profil utilisateur et les relations d'amitié dans les réseaux sociaux.

3. L'établissement et la standardisation de l'évaluation de résumé vidéo. La plupart de méthodes d'évaluation sont des questionnaires. Et il y a aucune évaluation automatique qui a réussi d'évaluer le résumé vidéo. Donc, il est difficile de comparer les algorithmes de résumé vidéo. Même si dans cette thèse on a proposé VERT, mais, cette mesure peut fonctionner pour un résumé forme, story-board. The meilleurs évaluation est de construire une mesure standard et automatique comme celle de TRECVID.

En se référant a ce travail, on peut améliorer l'algorithme de résumé le mesure d'évaluation en suivant ces aspects:

1. On peut introduire l'audio et le texte ensemble dans Video-MMR2 en mimant OB-MMR, ce qui peut aboutir des meilleurs résultats que OB-MMR.

2. Actuellement, l'ensemble optimisé des paramètres dans OB-MMR est générique, mais OB-MMR peut tirer d'avantages de paramètres basés sur le genre pour des vidéos ce diffèrent genres.

3. Pour OB-MMR, on peut utiliser seulement quelques traits sémantiques, mais il est possible d'exploiter quelques traits haut-niveaux traits sémantiques dans la vidéo, par exemple, des objets spécifiques.

4. VERT est convenable seulement pour le résumé stationnaire composé des images-clé. Donc, VERT ne peut pas traiter le résumé avec l'audio et le texte. Il est possible de étendre l'application de VERT en considérant l'audio ou le texte.

5. Il est intéressant d'utiliser une plateforme comme Amazon Mechanical Turk pour obtenir la vérité terrain depuis l'être humain pour affiner les poids pour des méthodes différentes et pour la sélection de VERT.

Bibliography

M. Barbieri, L. Agnihotri, and N. Dimitrova. Video summarization: methods and landscape. *Internet Multimedia Management Systems IV. Edited by Smith, John R. and Panchanathan, Sethuraman and Zhang, Tong. Proceedings of the SPIE*, November 2003.

R. Cai, L. Lu, H. Zhang, and L. Cai. Highlight sound effects detection in audio stream. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 37–40, Baltimore, MD, USA, July 2003.

J. Calic, D. Gibson, and N. Campbell. Efficient layout of comic-like video summaries. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(7):931–936, 2007.

J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR conference*, Melbourne Australia, August 1998.

H. S. Chang, S. Sull, and L. S. U. Efficient video indexing scheme for content-based retrieval. *IEEE Trans. Circ. Syst. Video Technol. 9, 8*, pages 1269–1279, 1999.

J. Chen, J. Xiao, and Y. Gao. iSlideshow: a Content-Aware Slideshow System. In *Proceedings of International Conference on Intelligent User Interfaces*, Hong Kong, China, February 7-10 2010.

M. G. Christel. Evaluation and user studies with respect to video. In *Multimedia Content Analysis, Management, and Retrieval*, San Jose, CA, USA, 2006.

F. Coldefy, P. Bouthemy, M. Betser, and G. Gravier. Tennis video abstraction from audio and visual cues. In *Proceedings of the 6th IEEE Workshop on Multimedia Signal Processing*, pages 163–166, Siena, Italy, 29 September-1 October 2004.

R. Cole, editor. *Survey of the state of the art in human language technology.* Cambridge University Press, New York, NY, USA, 1997. ISBN 0-521-59277-1.

M. Cooper and J. foote. Summarizing video using non-negative similarity matrix factorization. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing (MMSP)*, pages 25–28, St. Thomas, US Virgin Islands, 2002.

D. Das and A. F. Martins. A survey on automatic text summarization. Technical report, Literature Survey for the Language and Statistics II course at CMU, November 2007.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

N. Dimitrova. Context and memory in multimedia content analysis. *IEEE Multimedia 11*, pages 7–11, 2004.

N. D. Doulamis, A. D. Doulamis, Y. S. Avrithis, and S. D. Kollias. Video content representation using optimal extraction of frames and scenes. In *Proceedings of the ICIP Conference*, pages 875–879, Chicago, IL, 1998.

E. Dumont and B. Merialdo. Automatic evaluation method for rushes summary content. *International Workshop on Content-Based Multimedia Indexing*, pages 451–457, June 2008.

A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing 12 (7)*, pages 796–807, 2003.

B. Erol, D.-S. Lee, and J. Hull. Multimodal summarization of meeting recordings. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 3, pages 25–28, 2003.

M. Fayzullin, V. Subrahmanian, M. Albanese, and A. Picariello. The priority curve algorithm for video summarization. In *Proceedings of the 2nd ACM International Workshop on Multimedia Databases*, pages 28–35, Arlington, VA, USA, 13 November 2004.

A. M. Ferman and A. M. Tekalp. Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Transactions on Multimedia*, pages 244–256, 2003.

M. Furini and V. Ghini. An audio-video summarization scheme based on audio and video analysis. *Consumer Communications and Networking Conference*, 2006.

Y. Gao, W.-B. Wang, J.-H. Yong, and H.-J. Gu. Dynamic video summarization using two-level redundancy detection. *Springer Science + Business Media, LLC*, 2009.

R. Hammoud and R. Mohr. A probabilistic framework of selecting effective key frames for video browsing and indexing. In *Proceedings of the International Workshop on Real-Time Image Sequence Analysis*, 2000.

A. Hanjalic and M. Larson. Frontiers in multimedia search. In *Proceedings of ACM Multimedia*, Scottsdale, Arizona, USA, November 28Ű December 1 2011.

A. Hanjalic and H. Zhang. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Trans. Circ. Syst. Video Technolo*, 8: 1280–1289, Dec. 1999.

L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. In *Proceedings of the 7th ACM International Multimedia Conference*, pages 489–498, Orlando, FL, USA, 1999.

E. Hovy. Toward finely differentiated evaluation metrics for machine translation. *The Eagles Workshop on Standards and Evaluation*, 1999.

S. IMPEDOVO, L. OTTAVIANO, and S. OCCHINEGRO. Optical character recognition - a survey. *International Journal of Pattern Recognition and Artificial Intelligence*, 5, 1991.

IRISA. Spro toolkit. `http://www.irisa.fr/metiss/guig/spro`.

A. Jaimes, T. Echigo, M. Teraguchi, and F. Satoh. Learning personalized video highlights from detailed MPEG-7 metadata. In *Proceedings of the IEEE International Conference on Image Processing*, pages 133–136, New York, NY, USA, 22-25 September 2002.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks IV*, pages 1942–1948, 1995.

A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, New York, NY, USA, April 5-10 2008.

A. Komlodi and G. Marchionini. Key frame preview techniques for video browsing. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, pages 118–125. ACM Press, 1998.

M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: state of the art and challenges. *ACM Transactions on Multimedia Computing*, 2006.

Y. Li and B. Merialdo. Evaluation of video summaries. *8th International Workshop on Content-Based Multimedia Indexing*, pages 148–151, June 23-25 2010a.

Y. Li and B. Merialdo. Multi-video summarization based on AV-MMR. *8th International Workshop on Content-Based Multimedia Indexing*, pages 44–49, June 23-25 2010b.

Y. Li and B. Merialdo. VERT: A package for automatic evaluation of video summaries. *3rd International Workshop on Automated Information Extraction in Media Production*, October 25-29 2010c.

Y. Li and B. Merialdo. VERT: automatic evaluation of video summaries. *ACM Multimedia Conference*, October 25-29 2010d.

Y. Li and B. Merialdo. Multi-video summarization based on Video-MMR. *11th International Workshop on Image Analysis for Multimedia Interactive Services*, April 12-14 2010e.

Y. Li and B. Merialdo. Multi-video summarization based on OB-MMR. *9th International Workshop on Content-Based Multimedia Indexing*, 13-15 June 2011.

Y. Li and B. Merialdo. Multi-video summarization based on Balanced AV-MMR. *The 18th International Conference on MultiMedia Modeling*, 2012.

Y. Li, F. Wang, and B. Merialdo. Visualization of multi-video summaries demonstration. *7th International Workshop on Content-Based Multimedia Indexing*, June 3-5 2009.

Y. Li, B. Merialdo, M. Rouvier, and G. Linares. Static and dynamic video summaries. *ACM Multimedia Conference*, 28 November-1 December 2011.

R. Lienhart. Dynamic video summarization of home video. In *M.M. Yeung, B. Yeo, C.A. Bouman (Eds.), Storage and Retrieval for Media Databases: Proceedings of SPIE*, volume 3972, page 378Ű389, 2000.

R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Communications of the ACM 40 (12)*, pages 55–62, 1997.

C. Lin and B. Tseng. Optimizing user expectations for video semantic filtering and abstraction. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, volume 2, pages 1250–1253, Kobe, Japan, 2005.

C.-Y. Lin. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25-26 2004.

C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Human Technology Conference*, Edmonton, Canada, May 27 2003.

D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee, 1999.

D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

S. Lu, M. Lyu, and I. King. Video summarization by video structure analysis and graph optimization. In *In Proceedings of the International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, 2004a.

S. Lu, M. Lyu, and I. King. Video summarization by spatial-temporal graph optimization. In *Proceedings of the IEEE Symposium on Circuits and Systems (ISCAS)*, Vancouver, Canada, 2004b.

Y. Ma, X. Hua, L. Lu, and H. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia 7*, pages 907–919, 2005.

A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of visual communication and image representation*, April 2007.

S. Naci, U. Damnjanovic, B. Mansencal, J. Benois-Pineau, C. Kaes, M. Corvaglia, E. Rossi, and N. Aginako. The cost292 experimental framework for rushes summarization task in trecvid 2008. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 40–44. ACM, 2008.

M. Nilsson, J. Nordberg, and I. Claesson. Face detection using local smqt features and split up snow classifier. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

I. Otsuka, K. Nakane, A. Divakaran, K. Hatanaka, and M. Ogawa. A highlight scene detection and video summarization system using audio feature for a personal video recorder. *IEEE Transactions on Consumer Electronics 51 (1)*, 2005.

P. Over, A. Smeaton, and P. Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the international workshop on TRECVID video summarization*, pages 1–15. ACM, 2007.

P. Over, A. Smeaton, and G. Awad. The TRECVID 2008 BBC rushes summarization evaluation. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 1–20. ACM, 2008.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, July 2002.

R. Poli, J. Kennedy, and T. Blackwell. Particle swarm optimization: An overview. *Swarm Intell*, pages 33–57, 2007.

D. R. Radev, E. Hovy, and K. McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, pages 399–408, 2002.

R. Radhakrishnan, Z. Xiong, A. Divakaran, and T. Kan. Time series analysis and segmentation using eigenvectors for mining semantic audio label sequences. In *Proceedings of the International Conference on Multimedia and Expo*, 2004.

E. Rossi, S. Benini, R. Leonardi, B. Mansencal, and J. Benois-Pineau. Clustering of scene repeats for essential rushes preview. In *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS'09. 10th Workshop on*, pages 234–237. IEEE, 2009.

S. Rudinac, A. Hanjalic, and M. Larson. Finding representative and diverse community contributed images to create visual summaries of geographic areas. In *Proceedings of ACM Multimedia*, Scottsdale, Arizona, USA, November 28-December 1 2011.

Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the ACM Multimedia Conference (ACMMM)*, Los Angeles, CA, 2000.

K. Schoeffmann and L. Boeszoermenyi. Video browsing using interactive navigation summaries. In *Proceedings of Seventh International Workshop on Content-Based Multimedia Indexing*, Chania, Crete Island, Greece, June 3-5 2009.

H. Shih and C. Huang. MSN: statistical understanding of broadcasted baseball video using multi-level semantic network. *IEEE Transactions on Broadcasting 51*, pages 449–459, 2005.

F. Shipman, A. Girgensohn, and L. Wilcox. Creating navigable multilevel video summaries. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 753–756, Baltimore, MA, USA, 2003.

G. d. Silva, T. Yamasaki, and K. Aizawa. Evaluation of video summarization for a large number of cameras in ubiquitous home. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 820–828, Singapore, November 2005.

A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-495-2. doi: http://doi.acm.org/10.1145/1178677.1178722.

M. Sugano, Y. Nakajima, and H. Yanagihara. Automated MPEG audio-video summarization and description. *The International Conference on Image Processing*, 2002.

C. M. Taskiran. Evaluation of automatic video summarization systems. *Proceedings of SPIE*, 2006.

C. M. Taskiran, A. Amir, D. Ponceleon, and E. J. Delp. Automated video summarization using speech transcripts. In *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases*, pages 371–382, San Jose, CA, USA, 20-25 January 2002.

B. T. Truong and S. Venkatesh. Video abstraction: a systematic review and classification. *ACM Transaction Multimedia Compuatation Communication Application 3*, February 2007.

S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video manga: Generating semantically meaningful video summaries. In *Proceedings ACM Multimedia*, pages 383–392, Orlando, FL, October 30 1999.

University of Cambridge. HTK toolkit. `http://htk.eng.cam.ac.uk`.

Video Retrieval Group, City U. of Hong Kong. Local interest point extraction toolkit. `http://vireo.cs.cityu.edu.hk`.

H. D. Wactlar. Multi-document summarization and visualization in the informedia digital video library. In *Proc. of the 12th New Information Technology Conference*, Beijing, China, May 2001.

F. Wang and B. Merialdo. Multi-document video summarization. *International Conference on Multimedia and Expo*, June 2009.

C. Xu, X. Shao, N. C. Maddags, and M. S. Kankanhalli. Automatic music video summarization based on audio-visual-text analysis and alignment. *ACM SIGIR*, 2005.

M. Xu, C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 281–284, Baltimore, USA, 6-9 July 2003.

I. Yahiaoui, B. Merialdo, and B. Huet. Automatic video summarization. *Multimedia Content-based Indexing and Retrieval*, 2001.

B. Yu, W.-Y. Ma, K. Nahrstedt, and H.-J. Zhang. Video summarization based on user log enhanced link analysis. In *Proceedings of the 11th Annual ACM International Conference on Multimedia*, pages 382–391, Berkeley, CA, USA, November 2003.

D. Zhang and S.-F. Chang. Event detection in baseball video using superimposed caption recognition. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 315–318, Juan-les-Pins, France, November 2002.

F. Zheng, G. Zhang, and Z. Song. Comparison of different implementations of MFCC. *Journal on Computer Science and Technology*, pages 582–589, Sep 2001.

Index