Multi-Video Summarization Based on AV-MMR

Yingbo Li and Bernard Merialdo
Institut Eurecom
BP 193, 06904 Sophia-Antipolis, France
{Yingbo.Li, Bernard.Merialdo}@eurecom.fr

Abstract

This paper presents an algorithm for video summarization, Audio Video Maximal Marginal Relevance (AV-MMR), exploiting both audio and video information. It is an extension of the Video Maximal Marginal Relevance (Video-MMR) algorithm which was only based on visual information. AV-MMR iteratively selects segments which best represent unselected information and are non redundant with previously selected information. As for Video-MMR, AV-MMR is a generic algorithm which is suitable for both single and multiple videos with multiple genres. Several variants of AV-MMR are proposed and the best one is identified by experimentation. Besides, a visual representation of the coherence of audio and video information for a set of audio-visual sequences is also proposed.

1. Introduction

The amount of available multimedia data is growing daily. Among multimedia data, one may find personal digital videos, TV recordings, movies trailers and excerpts, or material from various other sources. Video summarization is one of the useful methods to browse these videos. Video summarization creates a short version of the original video, so that people can quickly understand the contents of the original video and easily make a decision about whether watching the whole content or not. Until now a lot of efforts have been devoted to the summarization of a single video sequence [1] [2], and more recently researchers have began focusing on the summarization of multiple videos [3] [4]. However, most of the research on video summarization considers only the video track of the audio-visual sequence, ignoring the information contained in the audio track. This can be a reasonable choice for certain types of videos, where the visual content is most important, but this is a limitation in the general case.

Some algorithms have been proposed [5] [6] [7] for summarization using both audio and video information. But these methods are often domain-specific, for example focusing on music video clips. In [5], the authors utilize motion features based on MPEG-7 and detect highlight by analyzing audio class and audio level. In [6] M. Furini removes some silent segments from original videos after detecting the silences in the audio. In music video summarization [7], the authors detect the chorus in audio and the repeated shots in video track. Generic approaches to video summarization are still an important issue.

In the domain of text summarization [8], Maximal Marginal Relevance (MMR) [9] has been proved to be a successful algorithm for text documents. Since text summarization and video summarization both aim at getting a shorter version of the original document, we have proposed the Video Maximal Marginal Relevance (Video-MMR) [4] to extend MMR into the domain of video summarization. Video-MMR is suitable for processing not only a single video, but also multi-video documents. In this paper we bring the audio information into the Video-MMR algorithm and propose a novel algorithm, Audio Video Maximal Marginal Relevance (AV-MMR). AV-MMR uses both the visual information (Bag of Visual Words [12]) and the audio information (Mel-frequency cepstral coefficients (MFCCs) [10] [11]). 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.

This paper is organized as follows: Section 2 will first review the principles of MMR and Video-MMR. Then Section 3 proposes the AV-MMR algorithm. Section 4 includes experimental results: a visualization method to demonstrate the coherence of audio and video features is presented in Subsection 4.1; Subsection 4.2 uses the method of summary reference comparison (SRC) to choose the best variant and parameters for AV-MMR; Video-MMR and AV-MMR

are compared in Subsection 4.3. Finally, Section 5 concludes this paper.

2. Related work

2.1. Text summarization and MMR

In the domain of Natural Language Processing [8], the research on text summarization has attracted a lot of attention. Text summaries are the condensed edition of the original text sets. Existing algorithms to perform text summarization for the single texts or multiple text sets include information fusion, graph spreading activation, centroid based summarization and multilingual multi-document summarization. Maximal Marginal Relevance (MMR) proposed by J. Carbonell and J. Goldstein [9] is a successful algorithm. MMR is based on the idea of Marginal Relevance (MR). MMR is an iterative process that incrementally builds a selection by adding elements one by one. MR uses two relations to select the best element at each step: one is the relation between this element and the intended content; the other is the relation between this element and the already selected elements. This idea is that the element selected should at the same time be representative of the intended content, but also different from the already selected elements to ensure diversity. For example, in the case of the selection of documents while answering a query, the MR of a document with respect to the query *Q* and the current selection S is defined by the equation:

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

Where Q is a query or user profile, and D_i and D_j are text documents in the returned list of documents *R* for the query Q. D_j is a document already selected in *S*, while D_i is a candidate in the list of unselected documents $R \backslash S$. In Eq. 1, the first term favors documents that are relevant to the topic, while the second could encourage documents containing new information which has not been selected yet. The parameter λ controls the proportion between query relevance and information novelty. Marginal Relevance can be applied to multi-document summarization 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:

$$D_{MMR} = arg\, max_{D_i \in R \backslash S} MR(D_i) \quad (2)$$

By iteratively selecting the text with largest MR in the text document or multi-documents, a text summary can easily be constructed.

2.2. Video summarization and Video-MMR

The goal of video summarization is to select the most important instants in a video or a set of videos. Because of the similarity between text summarization and video summarization, MMR is easily extended to the video domain as we have proposed in Video-MMR [4]. When iteratively selecting keyframes to construct a summary, Video-MMR selects a keyframe whose visual content is similar to the content of the videos, but at the same time different from the frames already selected in the summary. Video Marginal Relevance (Video-MR) may be defined as:

$$Video\text{-}MR(f) = \lambda\, Sim_1(f, V \backslash S) - (1-\lambda) \max_{g \in S} Sim_2(f, g) \quad (3)$$

where *V* is the set of all frames in all videos, *S* is the current set of selected frames, *g* is a frame in *S* and *f* 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):

$$S_{k+1} = S_k \cup \underset{f \in V \backslash S_k}{argmax} \left(\begin{matrix} \lambda\, Sim_1(f, V \backslash S_k) - \\ (1-\lambda) \max_{g \in S_k} Sim_2(f, g) \end{matrix} \right) \quad (4)$$

Sim_2 is just the similarity $sim(f, g)$ between frames f and g. We investigated two variants for $Sim_1(f, V \backslash S)$:

- The arithmetic mean of similarities:

$$AM(f, V \backslash S) = \frac{1}{|V \backslash (S \cup f)|} \sum_{g \in V \backslash (S \cup f)} sim(f, g) \quad (5)$$

- The geometric mean of similarities:

$$GM(f, V \backslash S) = \left[\prod_{g \in V \backslash (S \cup f)} sim(f, g) \right]^{\frac{1}{|V \backslash (S \cup f)|}} \quad (6)$$

The variants with AM and GM are named as AM-Video-MMR and GM-Video-MMR. The parameter λ allows adjusting the relative importance of relevance and novelty.

In [4], the authors compute the distance between each summary and the original videos for different combination of parameter λ and variants. The minimum distance means that the corresponding summary is the most similar with the original video, and is to be preferred. We will also exploit a similar comparison method for AV-MMR to decide which the best among variants is.

3. Audio Video Maximal Marginal Relevance

A video sequence contains both an audio and a video track. Here, we extend Video-MMR to Audio Video Maximal Marginal Relevance (AV-MMR) by considering information from both audio and video. We subsample the video sequence by extracting one

keyframe every 25 frames, so that one keyframe represents the visual content of one second sequence. To each keyframe, we associate the corresponding one second audio segment. We then modify Eq. 4. into Eq. 7. which defines how summary S_{k+1}can be constructed by iteratively selecting a new keyframe:

$$S_{k+1} = S_k \cup \underset{f \in V \setminus S_k}{argmax}[\lambda\, Sim_{I1}(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g) + \mu\, Sim_{A1}(f, V \setminus S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)] \quad (7)$$

where Sim_{I1} and Sim_{I2} are the same measures as Sim_1 and Sim_2 in Eq. 4. Sim_{A1} and Sim_{A2} play roles similar to Sim_{I1} and Sim_{I2} , but use the audio information of f. Eq. 7 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. 8.:

$$S_{k+1} = S_k \cup \underset{f, f' \in V \setminus S_k}{argmax}[\lambda\, Sim_{I1}(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g) + \mu\, Sim_{A1}(f', V \setminus S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f',g)] \quad (8)$$

The algorithm based on Eq. 8 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. 5. GM-AV-MMR uses the definition in Eq. 6. Parameter λ controls 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. 7 and Eq. 8, 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:

$$S_1 = \underset{f_i, f_i \neq f_j}{arg\ max} [\prod_{j=1}^{n} Sim_I(f_i, f_j) \cdot \prod_{j=1}^{n} Sim_A(f_i, f_j)]^{\frac{1}{n}} \quad (9)$$

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.

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

$$f_k = arg\ max_{f \in V \setminus S_{k-1}}[\lambda\, Sim_{I1}(f, V \setminus S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g) + \mu\, Sim_{A1}(f_i, V \setminus S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)] \quad (10)$$

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

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

In Section 4, 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.

4. Experimental results

This research is performed in cooperation with the news aggregator website (http://www.wikio.fr). This website collects videos from different sources, such as news, personal videos and sports, and present them as coherent articles. Each video set has a specific topic, like a film or a ceremony. We have downloaded 63 video sets from this site. 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.

The visual content of a keyframe is represented by the Bag-Of-Word feature. We first detect Local interest points (LIPs) in the image, based on the Difference of Gaussian and Laplacian of Gaussian. Then we compute SIFT descriptors for each point. The SIFT descriptors are clustered by k-means into 500 groups to compose a 500 visual word vocabulary. The BOW feature vector of a keyframe is the histogram of the number of visual words that appear in the keyframe. The processing software for this processing is obtained from [12]. The similarity between two keyframes $sim(f_i, f_j)$ is computed as the cosine similarity of the visual word histograms:

$$sim_I(f_i, f_j) = \cos\left(H_{f_i}, H_{f_j}\right) = \frac{H_{f_i} \cdot H_{f_j}}{\|H_{f_i}\| \|H_{f_j}\|} \quad (11)$$

where H_{f_i} and H_{f_j} are the visual word histograms of keyframes f_i and f_j.

To represent the audio content of a one second audio segment, we use the common Mel-frequency cepstral coefficients (MFCCs) [10]. The software to get MFCC vectors, SPro Toolkit, is from [11]. According to selected parameters, SPro creates 100

MFCC vectors per second, with 21 coefficients in each MFCC vector. We average these 100 MFCC vectors to obtain the audio feature vectors, S_{MFCC}. The similarity between two averaged MFCC vectors is computed and normalized as:

$$sim_A(a_i, a_j) = 1 - \frac{|a_i - a_j|}{\max_{a_m, a_n \in S_{MFCC}} (|a_m - a_n|)} \quad (12)$$

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 equation Eq. 13.

$$X' = \frac{X - \mu}{\sigma} \quad (13)$$

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.

Now, we first illustrate the coherence between audio and video within a video set. This illustration is useful to have a representation of the content of various videos inside a set. Then in Subsection 4.2, we select the best parameters and variant for AV-MMR, just as the authors did in [4]. Finally, we compare the experimental results of AV-MMR and Video-MMR in Subsection 4.3.

4.1. Circle representation of video and audio keyframes

Visualization of the content of a set of videos is a useful tool for understanding the possible relations between the various videos. In [13], the authors exploit a circle space to visualize the relations of keyframes from multi-video sets.

Assume that we have a set of videos, $\{V_1 \ldots V_m \ldots V_n\}$. We compute the similarity between a video V_m and a keyframe f as the maximum similarity value between frame f and another frame k of video V_m. This is represented by the following formula:

$$SIM(f, V_m) = \max_{k \in V_m, k \neq f} \{sim(f, k)\} \quad (14)$$

where k is a frame from video V_m.

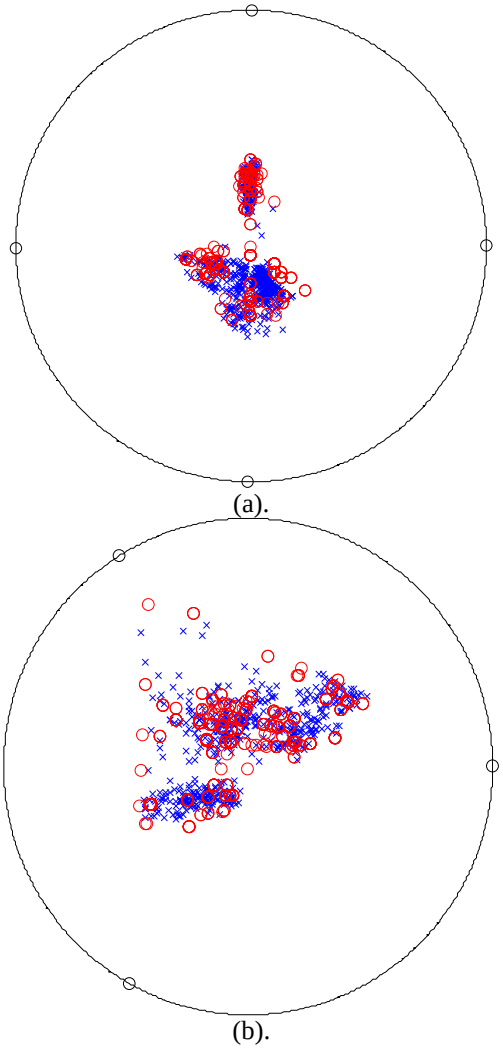
A circle space is used to visualize the relations between videos and frames. The videos are represented by points that are placed regularly on the circle boundary. Frames are represented by points inside the circle, with coordinates that are computed as:

$$P_f = \sum_{m=1}^{n} P_m \cdot SIM(f, V_m) \quad (15)$$

Where P_f is the position of the frame f, P_m is the position of video m. 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 close to the center of circle space, it means that this frame has an average similarity with all the videos.

Eq. 14 and 15 are suitable for both image and audio information. For image information $sim(f, k)$ is $sim'_I(f, k)$; while for audio information, $sim(f, k)$ is $sim'_A(f, k)$.

Two examples of circle representation for two different video sets are shown in Fig. 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. 1(a) and 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.



(a).

(b).

Fig.1. Circle representation: In the circle "×"=audio frame;"o"=video frame. On the circle, "o"=video.

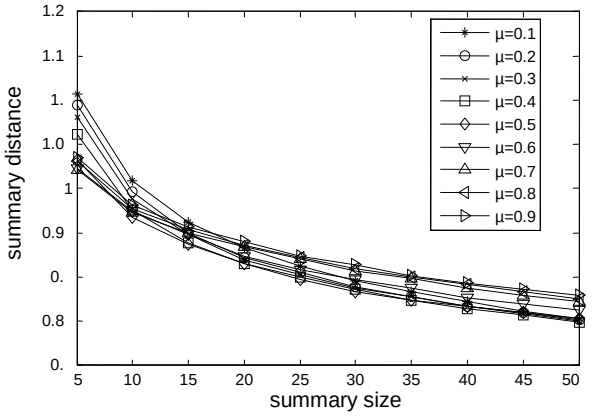
4.2. Summary reference comparison

In Video-MMR [4], the authors compare the results of different variants and parameters to select the best variant and parameter values. In this paper, we use the same approach to obtain the best variant and parameter values. The name of this method is Summary Reference Comparison (SRC). For this, we first define the distance between a video summary and the original video. We need to consider both the visual and the audio information, so we define the distance between a video set V and its summary S by the following equation:

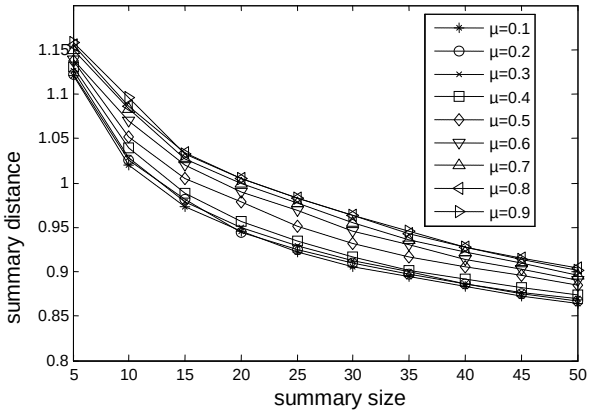
$$d(S,V) = \frac{1}{n}\sum_{j=1}^{n} \min_{f_j \in V, g \in S} \left[1 - \frac{sim_I'(f_j,g) + sim_A'(f_j,g)}{2}\right] \quad (16)$$

where n is the number of frames in 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. 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. We only consider SAV-MMR, and we compare the two variants, AM-AV-MMR and GM-AV-MMR.



(a) SRC of AM-AV-MMR



(b) SRC of GM-AV-MMR

Fig. 2. SRC of AV-MMR

Fig. 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. 2 are the mean distances over 62 video sets. The residual 63rd video set will be used to demonstrate the effect of AV-MMR in Section 4.3.

From these evaluations, we can see that the minimum summary distance is obtained in Fig. 2(a) with a value of $\mu = 0.5$. The corresponding value of the parameter λ is $\lambda = 0.7$. The values of AM-AV-MMR are globally smaller than GM-AV-MMR, so we can conclude that AM-AV-MMR generates better summaries.

4.3. 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", contains 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. 3.

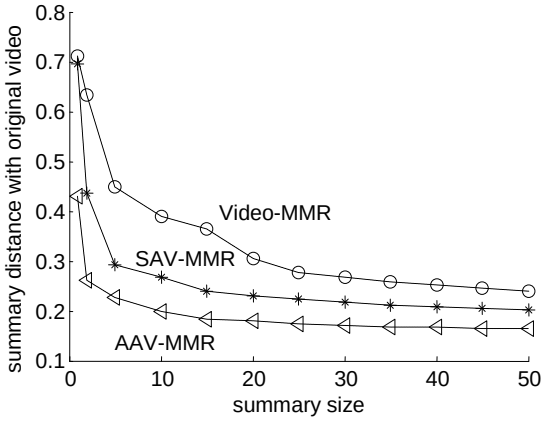


Fig.3. Comparison of Video-MMR and AV-MMR

From Fig. 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. Conclusions

In this paper, we have extended our previous 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 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. We have also shown that the AV-MMR algorithm produces better summaries than the previous Video-MMR. Finally, we have also proposed a visualization method to illustrate the audio-visual content of a set of videos.

Summarization of audio-visual material remains an important problem. Future work will be directed on more elaborate audio and visual processing, as well as new approaches for summary evaluation, taking into account human judgments.

6. Acknowledgements

This research is partially supported by the French ANR (Agence Nationale de la Recherche) in the project RPM2.

7. References

[1] Itheri Yahiaoui, Bernard Merialdo, Benoit Huet, "Automatic Video Summarization", *Multimedia Content-based Indexing and Retrieval*, Rocquencourt, France September 24 25, 2001.

[2] Arthur G. Money, Harry Agius, "Video summarisation: A conceptual framework and survey of the state of the art", *Journal on Visual Communication & Image Representation*, 121-143, 2008.

[3] Feng Wang and Bernard Merialdo, "Multi-document Video Summarization", *International Conference on Multimedia & Expo*, New York City, USA, 2009.

[4] Yingbo Li and Bernard Merialdo, "Multi-Video Summarization Based on Video-MMR", *International Workshop on Image Analysis for Multimedia Interactive Services*, Desenzano del Garda, Italy, 2010

[5] Masaru Sugano, Yasuyuki Nakajima, and Hiromasa Yanagihara, "Automated MPEG Audio-Video Summarization and Description", *Image Processing Proceedings International Conference on*, 2002.

[6] Marco Furini and Vittorio Ghini, "An Audio-Video Summarization Scheme Based on Audio and Video Analysis", *IEEE CCNC proceedings*, 2006.

[7] Changsheng Xu, Xi Shao, Namunu C. Maddags, Mohan S. Kankanhalli, "Automatic Music Video Summarization Based on Audio-Visual-Text Analysis and Alignment", *ACM SIGIR*, Salvador, Brazil August 15–19, 2005.

[8] Dipanjan Das Andr_e F.T. Martins, "A Survey on Automatic Text Summarization", *Literature Survey for the Language and Statistics II course at CMU*, November 2007.

[9] Jaime Carbonell and Jade Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", *ACM SIGIR conference*, Melbourne Australia, August 1998.

[10] Fang Zheng, Guoliang Zhang and Zhanjiang Song, "Comparison of Different Implementations of MFCC", *Journal on Computer Science & Technology*, 582-589, Sep. 2001.

[11] SPro Toolkit, *http://www.irisa.fr/metiss/guig/spro*

[12] *http://vireo.cs.cityu.edu.hk*. Video Retrieval Group, City University of Hong Kong.

[13] Yingbo Li, Feng Wang and Bernard Merialdo, "Visualization of Multi-Video Summaries", *7th International Workshop on Content-Based Multimedia Indexing*, Chania, Crete Island, Greece, June 3-5, 2009