Video Summarization Based on Balanced AV-MMR

Yingbo Li and Bernard Merialdo

EURECOM, Sophia Antipolis, France
{Yingbo.Li, Bernard.Merialdo}@eurecom.fr

Abstract. Among the techniques of video processing, video summarization is a promising approach to process the multimedia content. In this paper we present a novel summarization algorithm, Balanced Audio Video Maximal Marginal Relevance (Balanced AV-MMR or BAV-MMR), for multi-video summarization based on both audio and visual information. Balanced AV-MMR exploits the balance between audio information and visual information, and the balance of temporal information in different videos. Furthermore, audio genres and human face of each frame are analyzed in order to be exploited in Balanced AV-MMR. Compared with its predecessors, Video Maximal Marginal Relevance (Video-MMR) and Audio Video Maximal Marginal Relevance (AV-MMR), we design a novel mechanism to combine these indispensible features from video track and audio track and achieve better summaries.

Keywords: Multi-video summarization, MMR, Video-MMR, AV-MMR, Balanced AV-MMR.

1 Introduction

Nowadays video is ubiquitous in mobile phone, TV, Internet and so on. The amount of available videos is much more than the needs of a person, not mentioning that many videos are duplicates. Therefore the research on automatic video processing and retrieval has become a focused topic. Among the techniques for video processing, video summarization has been recognized as an important measure. For example, TRECVID [14] for rushes summarization has been an event in multimedia domain. Video summarization produces the summaries by analyzing the content of a source video stream, and condenses this content into an abbreviated descriptive form.

Currently many approaches of video summarization are proposed to process a single video [1] [2] [13], while there are many instances where multi-video data appears. For example, the YouTube website presents multiple related videos on the same webpage. Consequently, it is useful to discover the underlying relations inside a set of video. This need has been focused by some researchers and several successful algorithms [3] [4] have been developed. 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 existing algorithms [5] [6] consider both the audio and visual information in the summarization, but they are domain-specific. In [5] the authors consider that the video segments with silent audio are useless, but it is not always like that in the real videos. In [6], an algorithm has been proposed to summarize music videos: the chorus in audio track and the repeated shots in video track are detected. But the algorithms like [5] [6] only focus on a small domain, because in a domain-specific algorithm it is easier to utilize some special features or characteristics. For example, in sports video a loud ambient noise is a strong indication that the current visual information is likely to be important. However, in a generic algorithm we cannot rely on these specific characteristics. So there are not many generic algorithms exploiting both audio and visual information until now.

In this paper we propose a generic summarization algorithm by using both audio and visual information. Our algorithm is inspired by the previous algorithms, Video-MMR [4] by only using visual information, and AV-MMR in [12] by exploiting both audio and visual information. However, AV-MMR is a simple extension of Video-MMR, and does not consider the characteristics of audio and video track. Therefore, in this paper we propose a novel algorithm, Balanced AV-MMR, 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

The rest of the paper is organized as follows: Section 2 reviews the principles of Video-MMR and AV-MMR. Section 3 and Section 4 discuss the property of audio track and the importance of human face. And then the theory of Balanced AV-MMR is proposed to use the features in Section 5. After that in Section 6 we present the experimental results of Balanced AV-MMR. Our conclusion is in Section 7.

2 Review of Basic Systems

The series of Maximal Marginal Relevance (MMR) algorithms in video summarization originate from MMR algorithm in text summarization [11]. The first version is Video-MMR [4] based on pure visual information, which is extended to AV-MMR [12] by simply adding the audio part into the formula of Video-MMR. Before explaining our proposed algorithm, we need to first review these two MMR algorithms in video summarization.

2.1 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 [11] is easily extended to the video domain as Video-MMR in [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) is defined as:

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

where $Sim_1(f, V\backslash S) = \frac{1}{|V\backslash(S \cup f_i)|} \sum_{f_j \in V\backslash(S \cup f_i)} sim(f_i, f_j)$, 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-MMR:

$$S_{k+1} = S_k \cup \underset{f \in V\backslash S_k}{argmax} \left(\lambda\, Sim_1(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_2(f, g) \right) \qquad (2)$$

2.2 AV-MMR

In [12] the authors have proposed AV-MMR, an algorithm that exploits the information from both audio and video. Eq. 2 of Video-MMR is extended into Eq. 3.

$$S_{k+1} = S_k \cup \underset{f \in V\backslash S_k}{argmax} [\lambda\, Sim_{I1}(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f, g) + \\ \mu\, Sim_{A1}(f, A\backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f, g)] \qquad (3)$$

where Sim_{I1} and Sim_{I2} are the same measures as Sim_1 and Sim_2 in Eq. 2. Sim_{A1} and Sim_{A2} play roles similar to Sim_{I1} and Sim_{I2}. A and V are the collections of audio and video frames. Eq. 3 combines visual and audio similarities corresponding to the same frame, and it is called Synchronous AV-MMR.

3 Analysis of Audio Genres

Before exploiting audio information in Balanced AV-MMR it is necessary to analyze the characteristics of audio track. In this paper 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) [9], to construct a recognition system of audio genres for audio frames. In this paper 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 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 percentages 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 1, and Table 2 represents the percentages of audio segments. Since each category of video is analyzed separately, the sum of each row in Table 1 and Table 2 is 1.

Table 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

Table 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

In Table 1 and Table 2:

- *Sports* category obviously has the largest percentages 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.
- *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*, which is caused by the singing, even with music, being regarded as speech. Except above reason, another minor reason is that a few frames of speech are recognized as music because of our limited training data.

Limited to the space of paper, it is impossible to provide a complete description of the characteristics of every video category. 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 indispensible 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 paper, 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 existing AV-MMR.

4 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 [10] to detect the face in 89 video sets mentioned in Section 3. The percentage of frames with faces of each audio genre in all the frames is shown in Table 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 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.

Table 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 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. 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

Comparing Table 3 to Table 4, the difference is little so the influence of the frames comprising multiple faces is little. 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.
- 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:

- 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 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 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 of Section 3 and 4, 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 in 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 closer frames according to time seem to be redundant. Two frames in a video seem less different from 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.

5.1 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:

$$f_{k+1} = arg\ max_{f \in V \backslash S_k} \{\rho(f)[\lambda\, Sim_{I1}(f, V \backslash S_k)\ -\ (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g)]$$
$$+\big(1-\rho(f)\big)[\mu\, Sim_{A1}(f, A \backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)]\} \qquad (4)$$

In [12], it indicates $\lambda = 0.7$ and $\mu = 0.5$. Through bringing ρ into Eq. 4, 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. 4 is our fundamental formula for the following variants. When ρ is equal to 0.5, Eq. 4 degenerates into AV-MMR.

5.2 Balanced AV-MMR V1: using audio genre

Through the audio analysis in Section 3, 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 own 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, 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. $Sim_{A1}(f, A \backslash S_k)$ and $Sim_{A2}(f,g)$ in Eq. 4 become:

$$sim'_{A1}(f_i, A \backslash S_k) = \frac{1}{|A \backslash (S_k \cup f_i)|} \sum\nolimits_{f_j \in A \backslash (S_k \cup f_i)} \tau\big(f_i, f_j\big) sim\big(f_i, f_j\big)$$
$$sim'_{A2}(f,g) = \tau(f,g) sim(f,g) \qquad (5)$$

where $sim\big(f_i, f_j\big)$ and $sim(f,g)$ are original similarities by MFCC, same with the definitions in Eq. 3 and Eq. 4. And $\tau\big(f_i, f_g\big) = 1 + \theta_\tau \cdot (\theta_P - \big|P(f_i) - P(f_g)\big|)$. 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}:

$$\rho'(f) = \frac{\rho(f)}{\rho(f)+(1-\rho(f))\cdot(1+\varphi_{tr}(f))} = \frac{\rho(f)}{1+\varphi_{tr}(f)-\rho(f)\cdot\varphi_{tr}(f)} \quad (6)$$

Because of φ_{tr} and $\tau(f_i, f_j)$, the fundamental formula of Balanced AV-MMR, Eq. 4, transforms into the following formula:

$$\begin{aligned} f_{k+1} = arg\ max_{f\in V\backslash S_k}\{\rho'(f)[\lambda\, Sim_{I1}(f, V\backslash S_k) - \ (1-\lambda)\max_{g\in S_k} Sim_{I2}(f,g) \\ +\big(1-\rho'(f)\big)[\mu\, Sim'_{A1}(f, A\backslash S_k) - (1-\mu)\max_{g\in S_k} Sim'_{A2}(f,g)] \end{aligned} \quad (7)$$

5.3 Balanced AV-MMR V2: using face detection

According to the analysis in Section 4, 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 Subsection 5.2, it becomes:

$$\rho''(f) = \frac{\rho(f)\cdot\big(1+\beta_{face}(f)\big)}{\rho(f)\cdot\big(1+\beta_{face}(f)\big)+(1-\rho(f))\cdot(1+\varphi_{tr}(f))} = \frac{\rho(f)\cdot\big(1+\beta_{face}(f)\big)}{1+\varphi_{tr}(f)+\rho(f)\cdot(\beta_{face}(f)-\varphi_{tr}(f))} \quad (8)$$

where $\beta_{face}(f) = 1 + facenumber(f) * \theta_{face}$. 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. In semantic level, a frame comprising faces are more similar to another frame with faces than the frame without face. Also, two frames with the face often reveal the relevant content of the video, such as the different or same journalists in *News* and actors in *Movie*. Therefore the similarities Sim_{I1} and Sim_{I2} in Eq. 4 evolve into:

$$\begin{aligned} Sim'_{I1}(f_i, V\backslash S_k) &= \frac{1}{|V\backslash(S_k \cup f_i)|}\cdot \textstyle\sum_{f_j\in V\backslash(S_k\cup f_i)} \beta'_{face}\big(f_i, f_j\big) sim\big(f_i, f_j\big) \\ Sim'_{I2}(f,g) &= \beta'_{face}(f,g)\cdot sim(f,g) \end{aligned} \quad (9)$$

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

Based on above development, Eq. 7 of Balanced AV-MMR V1 can be reformulated as:

$$\begin{aligned} f_{k+1} = arg\ max_{f\in V\backslash S_k}\{\rho''(f)[\lambda\, Sim'_{I1}(f, V\backslash S_k) - (1-\lambda)\max_{g\in S_k} Sim'_{I1}(f,g)] \\ +\big(1-\rho''(f)\big)[\mu\, Sim'_{A1}(f, A\backslash S_k) - (1-\mu)\max_{g\in S_k} Sim'_{A2}(f,g)]\} \end{aligned} \quad (10)$$

5.4 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:

- Closer frames according to time in a video commonly represent more relevant content, so two closer frames in a video are regarded more similar than two further frames.
- 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

$$\alpha_{time}(f_i, f_j) = \begin{cases} 1 & , if\ f_i\ and\ f_j\ are\ from\ two\ videos; \\ 1 + \theta_{time} \cdot \left(1 - \frac{|t(f_i) - t(f_j)|}{10 * D_M}\right) & , if\ f_i\ and\ f_j\ are\ from\ the\ same\ video. \end{cases} \quad (11)$$

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:

$$Sim''_{I1}(f_i, V\backslash S_k) = \frac{1}{|V\backslash(S_k \cup f_i)|} \cdot \sum_{f_j \in V\backslash(S_k \cup f_i)} \beta'_{face}(f_i, f_j)\alpha_{time}(f_i, f_j)sim(f_i, f_j)$$

$$Sim''_{A1}(f_i, A\backslash S_k) = \frac{1}{|A\backslash(S_k \cup f_i)|} \cdot \sum_{f_j \in A\backslash(S_k \cup f_i)} \tau(f_i, f_j)\alpha_{time}(f_i, f_j)sim(f_i, f_j) \quad (12)$$

Sim'_{I2} and Sim'_{A2} are similarly multiplied by α_{time} and become Sim''_{I2} and Sim''_{A2}. Consequently, the formula of Balanced AV-MMR V3 is similar to Eq. 10 of Balanced AV-MMR V2 and generalized as

$$f_{k+1} = arg\ max_{f \in V\backslash S_k}\{\rho''(f)[\lambda\, Sim''_{I1}(f, V\backslash S_k) - (1-\lambda)\max_{g \in S_k} Sim''_{I1}(f, g)] + (1 - \rho''(f))[\mu\, Sim''_{A1}(f, A\backslash S_k) - (1-\mu)\max_{g \in S_k} Sim''_{A2}(f, g)]\} (13)$$

5.5 The procedure of Balanced AV-MMR

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

1) Detect the audio genres of the frames by HTK audio system described in Section 3, and the face by the toolkit in Section 4;
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:

$$S_1 = arg \max_{f_i, f_i \neq f_j} [\prod_{j=1}^{n} Sim_I(f_i, f_j) \prod_{j=1}^{n} Sim_A(f_i, f_j)]^{\frac{1}{n}} \quad (14)$$

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 visual information between f_i and f_j; while Sim_A is the similarity of audio;

4) Select the frame f_{k+1} by the formula of a variant of Balanced AV-MMR;
5) Set $S_{k+1} = S_k \cup \{f_{k+1}\}$.
6) Iterate to step 5) until S has reached the predefined size.

6 Experimental Results

Our experimental videos are 36 video sets from 7 categories mentioned in Section 3, 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 as $sim_I(f_i, f_j) = \cos\left(H_{f_i}, H_{f_j}\right) = \frac{H_{f_i} \cdot H_{f_j}}{\|H_{f_i}\|\|H_{f_j}\|}$, 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 [8]. The similarity of two averaged MFCC vectors is computed and normalized as $sim_A(a_i, a_j) = 1 - \frac{|a_i - a_j|}{\max\limits_{a_m, a_n \in S_{MFCC}}(|a_m - a_n|)}$, 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 $d_{AVD}(S, V) = \frac{1}{n}\sum_{j=1}^{n} \min\limits_{f_j \in V, g \in S}\left[1 - (sim_I(f_j, g) + sim_A(f_j, g))/2\right]$, 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 $d_{VD}(S, V) = \frac{1}{n}\sum_{j=1}^{n} \min\limits_{f_j \in V, g \in S}(1 - sim_I(f_j, g))$.

In the fundamental formula of BAV-MMR, Eq. 4, we need to decide the value of parameter ρ. In this paper we consider ρ as a constant value for different frames. To remain consistent with Video-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. 1. SRC here uses $d_{AVD}(S, V)$.

From Fig. 1 when $\rho = 0$, d_{AVD} is large when the summary size is small and vice versa. Since we want a commonly small d_{AVD} 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. 6 $\varphi_{tr}(f) = 0.1$ 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.

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. 2 and Fig. 3. We have not drawn the curve of the fundamental BAV-MMR with $\rho = 0.5$, which is the same with AV-MMR in Fig. 1. 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 improves audio track and video track in the summarization.

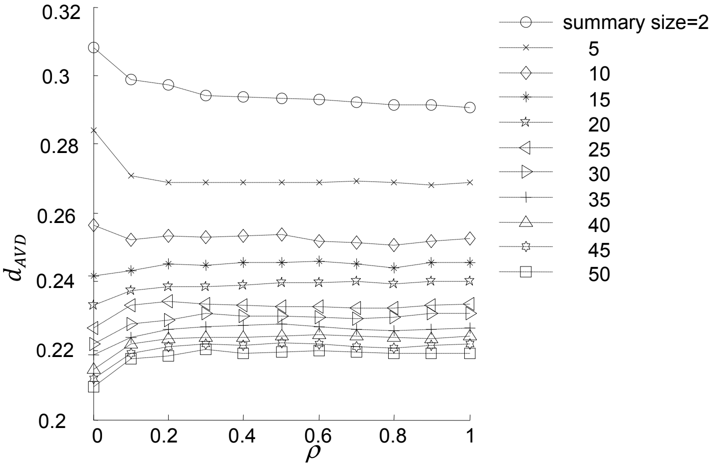
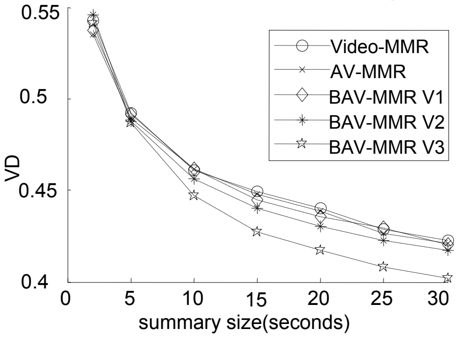


Fig. 1. SRC of ρ_T



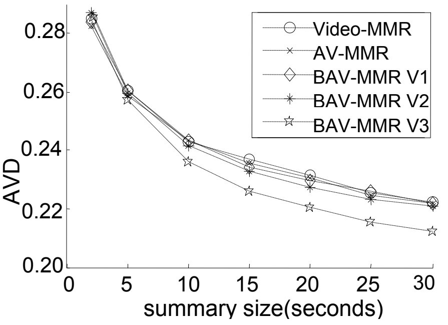


Fig. 2. VDs of different measures

Fig. 3. AVDs of different measures

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 the 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.

7 Conclusion

In this paper, we have proposed a novel 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 the weights in Balanced AV-MMR are manually decided, so it is necessary to automatically optimize the weights to the category of the video, summary size, and so on in the future.

References

[1] I. Yahiaoui, B. Merialdo, B. Huet, "Automatic Video Summarization", Multimedia Content-based Indexing and Retrieval, Rocquencourt, France, September 2001.

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

[3] F. Wang and B. Merialdo, "Multi-document Video Summarization", ICME, New York City, USA, 2009.

[4] Y. Li and B. Merialdo, "Multi-Video Summarization Based on Video-MMR, International Workshop on Image Analysis for Multimedia Interactive Services, Italy, 2010.

[5] M. Furini and V. Ghini, "An Audio-Video Summarization Scheme Based on Audio and Video Analysis", *IEEE CCNC proceedings*, 2006.

[6] C. Xu, X. Shao, N. C. Maddags, M. S. Kankanhalli. , "Automatic Music Video Summarization Based on Audio-Visual-Text Analysis and Alignment", ACM SIGIR, Salvador, Brazil, 2005.

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

[8] SPro Toolkit. http://www.irisa.fr/metiss/guig/spro

[9] http://htk.eng.cam.ac.uk. University of Cambridge.

[10] M. Nilsson, J. Nordberg, I. Claesson, "Face Detection using Local SMQT Features and Split Up SNoW Classifier", IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007.

[11] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", SIGIR, Melbourne Australia, 1998.

[12] Y. Li, B. Merialdo, "Multi-Video summarization based on AV-MMR", International Workshop on Content-based Multimedia Indexing, France, 2010.

[13] B. Truong and S. Venkatesh. "Video abstraction: A systematic review and classification", *ACM Transactions on Multimedia Computing*, *Communications and Applications*, 3(1), Jan 2007.

[14] TRECVID: http://trecvid.nist.gov/.