Noname manuscript No.
(will be inserted by the editor)

Multimedia Maximal Marginal Relevance for Multi-video Summarization

Yingbo Li · Bernard Merialdo

Received: date / Accepted: date

Abstract In this paper we propose several novel algorithms for multi-video summarization. The first and basic algorithm, Video Maximal Marginal Relevance (Video-MMR), mimics the principle of a classical algorithm of text summarization, Maximal Marginal Relevance (MMR). Video-MMR rewards relevant keyframes and penalizes redundant keyframes, only relying on visual features. We extend Video-MMR to Audio Video Maximal Marginal Relevance (AV-MMR) by exploiting audio features. Consequently, we also propose Balanced AV-MMR, which exploits additional semantic features, the balance between audio information and visual information, and the balance of temporal information in different videos of a set. The proposed algorithms are generic for various video genres, designed to summarize multi-video and using multimodal information in the video. Our series of MMR algorithms in video summarization are proved to be effective for summarizing multi-video by large-scale experiments.

Keywords Video summarization · Video summaries · MMR · Video-MMR · AV-MMR · Balanced AV-MMR

1 Introduction

The rapid increase of the amount of videos 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 technique to deal with video data. Research into video summarization produces an abbreviated form by extracting the most important and pertinent content in the video. The forms of video summaries can be categories into keyframes, being representative images of the source video, and video skims, being a collection of video segments much shorter than the source video [6]. Video summaries

Yingbo Li
EURECOM
E-mail: yingbo.li@eurecom.fr

Bernard Merialdo
EURECOM
E-mail: bernard.merialdo@eurecom.fr

can be used in various applications, such as searching systems and interactive browsing, which facilitates the users' demand of managing and accessing digital video content [26] [3] [34] [22] [1].

Many current summarization algorithms only consider the features from the video track and neglect the audio track or independently investigate the visual and audio information because it is hard to fuse audio information into the processing of visual information. In the summarization category by only exploiting visual information, visual attention model is an outstanding one. In [13] the authors propose an algorithm by detecting temporal gradient based dynamic visual saliency. The current popular approach of feature extraction, sparse representation [6] [21] [16], is also introduced into video summarization. The authors in [20] propose to use sparse representation to analyze the spatio-temporal information to extract keyframes from unstructured consumer videos without shot detection and semantic understanding. Z. Wang et al. [47] propose a sequence-kernel based sparse representation by constructing an optimal combination of the clustered dictionary. S. Rudinac et al [38] propose a novel visual summarization algorithm by approaching the large-scale human image selection obtained on the crowdsourcing platform of Amazon Mechanical Turk.

Several current algorithms [41] [15] [48] consider both audio track and video track, but they are domain specific. MPEG-7 motion activity descriptor and highlight detection by analyzing audio class and audio level are the main measures in [41]. In [15] the authors consider that the video segments corresponding to silence in the audio track are useless. In [48], the authors have invented an algorithm to summarize music videos: the authors detect the chorus in audio and the repeated shots in video track. The three successful algorithms above are examples of video summarization using both visual and audio information, but each of them only focuses on one domain. Many approaches, which summarize the video by both visual and audio information simultaneously, are domain specific. The reason is 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. However, some summarization algorithms exploiting both audio and visual information still exist. Visual attention model [29] is a classical summarization algorithm, which individually builds the attentions to the audio track and video track and then fuses the attentions to audio and video tracks to summarize the video. Another example of independent summarizations of video and audio tracks is that W. Jiang et al. [19] individually implements video summarization by image quality and face information, and audio summarization by audio genres, which conform to the high-level semantic rules. Topic-Oriented Multimedia Summarization [10] is proposed by fusing text, audio including speech and visual features related to specific topics. [28] is another summarization algorithm by detecting the high-level features influencing the emotion, such as the face and the music.

Besides the information inside the video itself, some approaches exploits the information outside the video like human actions when watching the video. For example, in [37] the authors propose to use users' viewing behaviors, such as eye movement, blink, and head motion, to measure the interesting video parts and construct video summary.

While a lot of efforts have been devoted to the summarization of a single video [34] [49] as most of the successful approaches mentioned above, less attention has been given to the summarization of a set of videos [49]. 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. Since the video is multimodal with the information of sound, music, still images, moving images and text [9], multi-video summarization is more complex than text summarization and other text processing techniques. In addition to the low-level features, multi-video summarization needs to consider semantics in the video to facilitate the understanding of the summary. Multi-video summarization has been studied by some researchers [45] [12] [8] [46] [7]. However, a non domain-specific multi-video summarization by using multimodal information inside the video is still an open problem worth to be focused on.

Our target is to propose a generic system of video summarization algorithms, which is suitable for summarizing multi-video of any genre (Documentary, News, Music, Advertisement, Cartoon, Movie, and Sports), by using multimodal information in the video, such as visual information, human face, acoustic information and so on. So our proposed algorithms are of the following properties together: generic, multi-video and multimodal.

1. Generic property. The proposed algorithms are able to summarize the videos of different genres: music, sports, advertisements, news and so on. Therefore, we don't need to know the genre of the video or a set of video, even the genres of the videos in a set could be various. So we try to optimize parameters in our algorithms for generic purpose.
2. Multi-video property. Our system could process one or multiple videos at the same time, and consider temporal character of different videos in a video set. In the proposed system we select the summary frames from the set of all the video frames to construct our summary, as most of the state-of-the-art systems. However, we significantly build in the semantic relation between inter-video and intra-video frames. The visual change between frames is commonly exploited. However, the acoustic change, especially the genre change of the audio segment, has not yet been considered in the state of the art. The change of the audio genre between the audio segment in a video is an obvious indication of the semantic transformation, while it does not exist between the frames of different videos. Furthermore, the semantic difference between videos is normally so great that it is not the same order of magnitude as the semantic difference between frames in one video. Even two frames far temporally from each other or near each other in a video should be of different semantic similarity. We exploit above semantic information, which cannot be replaced by simply setting some weight in inter-video and intra-video without considering video's property in the semantic level.
3. Multimodal property. In our system, we focus on the obtrusively sourced information [34], which means the information directly obtained from the video itself: visual features including face and others, audio features including audio genre, speech and others, and the possible text features including the text from speech and the text from the built-in video frames. Since the proposed system is generic, the available video features would be limited. The domain-specific features, for example, at the time of shooting in sports video, cannot be exploited.

Besides above contributions to the domain of video summarization, we also propose the semantic combination of multimedia features considering the video properties. In the current state of the art, it is popular to simply sum the weights of the features and summarize the video as the scenes with the highest weights, like [30]. It is a reasonable way to combine the features, but the semantic meanings lying under the features are ignored and not exploited. In the multimedia variants of our basic approach based on the video feature, we semantically consider the underlying relations between audio features and visual features, and semantic relations inside a video and between videos. We are not only considering the relations between the intrinsic features of the video, but also bringing the factor of the users' attention, the extrinsic factor, into the relations between intrinsic features, because the video is finally watched by users.

The content of this paper is the following: video summarization comes into the research community following the text summarization when the video becomes popular in our world. Maximal Marginal Relevance (MMR) [4] is a successful algorithm in text summarization by selecting the most important and common text words, so we borrow the idea of MMR into video summarization and propose Video-MMR by only exploiting visual features. But the principle of Video-MMR is not exactly the same as MMR. MMR constructs the summary depending on a static query, while Video-MMR dynamically constructs video summary depending on the summary under construction.

After we bring and adapt MMR into Video-MMR, we propose to extend Video-MMR by adding more multimedia information at the feature and semantic levels. We exploit audio information in Video-MMR and develop it to AV-MMR (Audio Video MMR), and Balanced AV-MMR step by step. AV-MMR simply extends Video-MMR. While, Balanced AV-MMR further considers some important semantic features in visual information and audio information, including the audio genre, the face, and the temporal feature, which is especially important for multi-video summarization, of videos in the same set. By considering temporal relations between video, Balanced AV-MMR better distinguishes multi-video summarization from the summarization of multiple video frames. The proposed algorithms are assessed by humans.

In addition, we make a large-scale analysis to the influence of video genres on video summarization, which is important for a robust algorithm but not implemented by many previous approaches.

This paper is organized as follows: In Section 2 we review the principle of MMR in text summarization. In Section 3.1 we introduce the MMR method into video summarization and propose Video-MMR. Then we improve our Video-MMR by adding acoustic cue and propose AV-MMR, and Balanced AV-MMR in the remaining paragraphs of Section 3. Then in Section 4 we compare the summaries, also with the ground truth, make a large-scale analysis of the summarization algorithms, deeply analyze the experimental results and suggest the best algorithm for the practical application. At last we present some conclusions in Section 5.

2 MMR in text summarization

Text summarization is a popular research topic in the area of Natural Language Processing [8] [33] [27]. Text summaries preserve important information, and are short compared with original single document or multiple documents. Since 1990s, a lot of work has dedicated to the research of text summarization algorithms for multiple documents [8] [32]. Various approaches have been proposed, such as information fusion, graph spreading activation, centroid based summarization and multilingual multi-document summarization. A popular and efficient algorithm of multi-document text summarization is MMR proposed by [4]. The Marginal Relevance (MR) of a document D_i with respect to a query Q and a document selection S is defined by the equation:

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

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:

$$D_{MMR} = argmax_{D_i \in R \backslash S} MR(D_i) \tag{2}$$

In [4], 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 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 Multimedia MMR algorithms

In the previous section, we reviewed the principle of MMR, which has been a successful algorithm in text summarization. Though text summarization and video summarization are not exactly the same, the tasks of both are to extract important information from a set of data. Consequently we propose to adapt MMR and introduce it into the domain of video summarization. In this section, we will first propose Video-MMR which only exploits visual features in the video. Then several multimedia MMR algorithms, AV-MMR, and Balanced AV-MMR which exploit both visual and acoustic features, are proposed.

3.1 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. So the forms of video summary [42] include stationary images, also called keyframes and storyboard, and moving images, also called video skims. Both forms of video summaries have their own advantages: keyframes are easy to display in a static space and easier to be understood, while video skims can show a lot of dynamic content together with audio segments. In this paper we focus on the selection of salient keyframes. We subsample the video at the rate of one frame per second. In [5] the authors select one frame per half second as the subsampling rate. While in our case multi-video is normally of a long duration, so the candidate frames from the subsampling of one frame per second is big enough for a summary with the size from 10 even to 50 keyframes. More frames per second would require more processing time, especially for the longer multi-video case. So one frame per second is a trade-off between the candidate contents and processing time. Furthermore, the relation between the visual contents in the summary, S, and in the original video, V, can be measured by the following similarity:

$$Sim(S, V) = 1 - \frac{1}{n}\sum_{j=1}^{n} \min_{f_j \in V, g \in S} d(f_j, g) \tag{3}$$

where n is the number of frames in V. g and f_j are frames respectively from S and V. And $d(f_j, g)$ is the distance normalized in $[0, 1]$. With this presentation, the best summary $\hat{S}$ (for a given length) is the one that achieves the maximum similarity:

$$\hat{S} = argmax_S[Sim(S, V)] \tag{4}$$

Because the principle of video summarization is similar to text summarization, we propose to adapt the MMR criteria to design a new algorithm, Video Maximal Marginal Relevance (Video-MMR) [23], 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, as illustrated in Fig. 1. By analogy with the MMR algorithm, we define Video Marginal Relevance (Video-MR) by:

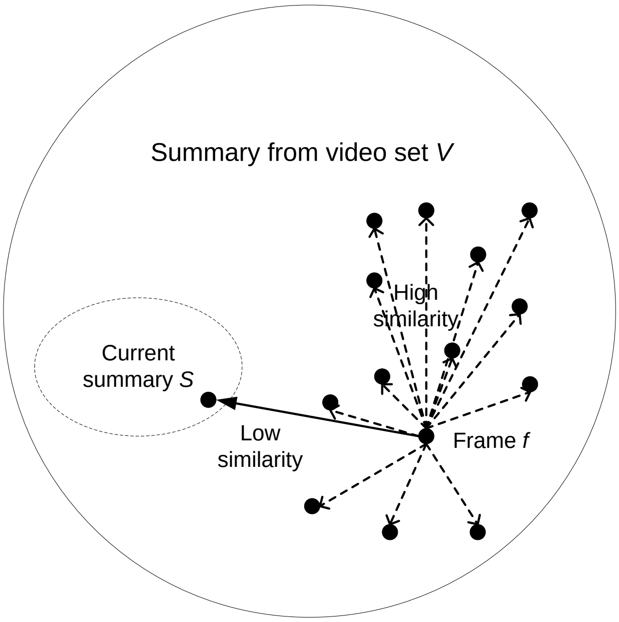

Fig. 1: The illustration of Video-MMR

$$\textit{Video-MR}_S(f) = \lambda Sim(f, V \backslash S) - (1-\lambda) Sim(f, S) \tag{5}$$

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. $V \backslash S$ represents the frames in V but not yet selected in S. Sim is the similarity between the video frames or video set. Based on this measure, a summary S_{k+1} can be constructed by iteratively selecting the keyframe with Video Maximal Marginal Relevance (Video-MMR):

$$S_{k+1} = S_k \bigcup argmax_{f \in V \backslash S_k} \{\lambda Sim_1(f, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_2(f, g)\} \tag{6}$$

We define Sim_1 as average frame similarity:

$$Sim_1(f_i, V \backslash S_k) = \frac{1}{|V \backslash (S_k \bigcup f_i)|} \sum_{f_j \in V \backslash (S_k \bigcup f_i)} Sim(f_i, f_j) \tag{7}$$

while $Sim_2(f, S_k) = \max_{g \in S_k} Sim_2(f, g)$ and $Sim_2(f, g)$ is just the similarity $Sim(f, g)$ between frames f_i and g. The parameter λ is used to adjust the relative importance of relevance and novelty. Alternatively, the formula of Video-MMR can be rewritten as:

$$S_{k+1} = S_k \bigcup argmax_{f \in V \backslash S_k} \{\lambda Sim_1(f, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_2(f, g)\} \tag{8}$$

while Sim_1 in MMR is the similarity to the static text query, but in Video-MMR it is the similarity to video summary which is dynamically constructed and a part of the source video. Assuming that the frame numbers of multi-video V and the desired summary size were separately N and K, time complexity of Video-MMR is $O(K^2N)$ same with MMR [14].

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:

$$f_1 = argmax_{f_i}(\prod_{j=1, f_i \neq f_j}^{n} Sim(f_i, f_j))^{1/n} \tag{9}$$

   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.
2. Select the frame f_k by Video-MMR:

$$f_{k+1} = argmax_{f_i \in V \backslash S_k}[\lambda Sim_1(f_i, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_2(f_i, g)] \tag{10}$$

3. Set $S_{k+1} = S_k \bigcup \{f_{k+1}\}$.
4. Iterate to step 2 until S has reached the desired size.

We describe the above steps as the following pseudocode for a clear description:

```
Initialize:
$S_1 = argmax_{f_i}(\prod_{j=1,f_i \neq f_j}^{n} Sim(f_i, f_j))^{1/n}$;
repeat
    $S_{k+1} = S_k \bigcup \{argmax_{f_i \in V \backslash S_k}[\lambda Sim_1(f_i, V \backslash S_k) - (1-\lambda)\max_{g \in S_k} Sim_2(f_i, g)]\}$
until $|S_{k+1}| = size_{desired}$;
```

Algorithm 1: The pseudocode of Video-MMR

The sparse representation based algorithms are highly focused nowadays and introduced into the domain of video summarization in the recent several years. The summarization algorithm of sequence-kernel based sparse representation [47] is a successful one in these algorithms. It resolves three problems of a good dictionary, measuring the reconstruction error and the optimal combination from the dictionary during the summarization procedure. While, if we regards the whole set of candidate video frames as the dictionary elements and each dictionary element is one video frame by mimicking the principle in [47], our proposed algorithm could be regarded similar to the principle in it. In Eq. 6 the summary is only a small part of frames in the video V, so the video information in the summary S is only a little compared to the total information in V and $V \backslash S$ can represent most information in V. When $\lambda = 1$, $Sim_1(f, V \backslash S)$ in Eq. 6 measures the reconstruction error of one dictionary element because the information inside current summary $V \backslash S$ contains most information of V. Thus the procedure of Video-MMR summarization is exactly the search for the optimal combination of the dictionary elements. Therefore, our proposed algorithm, Video-MMR can be considered as a variant of the algorithm of sequence-kernel based sparse representation in some sense.

3.2 AV-MMR

A video sequence contains both audio and video tracks. Consequently, we extend Video-MMR to Audio Video Maximal Marginal Relevance (AV-MMR) by considering information from both audio and video tracks. We associate the corresponding one second audio segment to each video frame. Then we can modify Eq. 6 into Eq. 11, which defines how summary S_{k+1} can be constructed by iteratively selecting a new keyframe:

$$\begin{aligned}
&S_{k+1} = S_k \bigcup argmax_{f \in V \backslash S_k} \\
&[\lambda Sim_{I1}(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g) + \\
&\mu Sim_{A1}(f, V\backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)] \\
&= S_k \bigcup argmax_{f \in V \backslash S_k} [MR_I(f, S_k) + MR_A(f, S_k)]
\end{aligned} \tag{11}$$

where Sim_{I1} and Sim_{I2} are the same measures as Sim_1 and Sim_2 in Eq. 6. Sim_{A1} and Sim_{A2} play roles similar to Sim_{I1} and Sim_{I2}, but use the audio information of f. To simplify the formula, we introduce the donation of MR_I and MR_A in the following sections. $MR_I(f, S_k) = \lambda Sim_{I1}(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g)$ and $MR_R(f, S_k) = \mu Sim_{A1}(f, V\backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)$.

AV-MMR also considers sim_{I1} and sim_{A1} as arithmetic mean average. And the parameter μ plays the same role with λ for audio information. By Eq. 11 we can construct AV-MMR summarization procedure like Video-MMR in Section 3.1. As well, the complexities of AV-MMR and the following AV-MMR based algorithms are also $O(K^2N)$ like Video-MMR.

3.3 Balanced AV-MMR

The study on the human attention suggests that in a short period (one second, for example) a person's attention is limited so that he cannot catch the overload information. [39] [50] [2] [11] [17] [31]. If the audio attracted more attention from the user, the user naturally and reasonably pays less attention to video content and vice versa. Therefore, the attention on audio information and visual information should be balanced for a video segment. Consequently we give our novel algorithm the name "balance". In this section, we will introduce the factors of audio genre, the face and the time to AV-MMR and propose the variants of Balanced AV-MMR, or called BAV-MMR, which improve the balance and the similarity of frames in the semantic level. We linearly introduce three weights of audio genre, the face and the time to adjust the balance, the visual similarity, and the audio similarity, because the linear weighting is a simple and straightforward way to enhance the influence from these factors.

3.3.1 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:

$$f_{k+1} = argmax_{f \in V \backslash S_k} \{\rho(f) \cdot MR_I(f, S_k) + (1-\rho(f)) \cdot MR_A(f, S_k)\} \tag{12}$$

Balanced AV-MMR considers the balance between audio and video by the weight ρ. When ρ increases, the visual information takes a more important role in Balanced AV-MMR, and vice versa. Eq. 12 is our fundamental formula for the following variants. When ρ is equal to 0.5, Eq. 12 degenerates into AV-MMR.

3.3.2 Balanced AV-MMR V1 (using audio genre)

Same with AV-MMR, here we also use one second audio segments corresponding to the keyframes as the unit for audio analysis. According to the analysis in [25], audio genre, being speech, music and silence in this paper, is an important feature in the video. Audio genre can influence the similarities between frames at the semantic level. For audio track two frames from the same audio genre are more similar than the similarity from two different genres. It is obvious that the audio frames with the same genre are more similar than the audio frames with different genres, when they have the same similarity according to the audio features. Here we give an example to better describe this assumption: when two pairs of audio segments, speech-speech pair and speech-music pair have the same similarity values according to low-level audio features, people will favor the speech-speech pair as more similar. No matter the genre of the video, such as Sports, TV and so on, it is impossible for the human to feel that speech-music pair is more similar than speech-speech pair.

But the similarity in semantic level cannot be reflected by the low-level feature. Consequently, we can introduce an augment factor for audio genres to adjust the similarity of audio features. Here we use τ to denote this factor and linearly adjust the similarity of audio features as $\tau \cdot sim(f_i, f_j)$. Eq. 12 and its $Sim_{A1}(f, A\backslash S_k)$ and $Sim_{A2}(f, g)$ becomes:

$$f_{k+1} = argmax_{f \in V \backslash S_k} \{\rho(f) \cdot MR_I(f, S_k) + (1 - \rho(f)) \cdot MR'_A(f, S_k)\} \tag{13}$$

where

$$\begin{aligned} &MR'_A(f, S_k) = (1 - \rho(f))[\mu Sim'_{A1}(f, A\backslash S_k) - (1 - \mu) \max\nolimits_{g \in S_k} Sim'_{A2}(f, g)]; \\ &Sim'_{A1}(f_i, A\backslash S_k) = \tfrac{1}{|A\backslash(S_k \bigcup f_i)|} \textstyle\sum_{f_j \in A\backslash(S_k \bigcup f_i)} \tau(f_i, f_j) Sim(f_i, f_j); \\ &Sim'_{A2}(f, g) = \tau(f, g) Sim(f, g) \end{aligned} \tag{14}$$

and $Sim(f_i, f_j)$ and $Sim(f, g)$ are original similarities of the audio, same with the definitions in Eq. 12. 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 genre, because the speech in the video attracts the human attention most compared to the other two kinds. These weights are decided manually in the experiment to slightly adjust the similarity according to audio genres.

Moreover, 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. So 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. ρ in Eq. 13 displays the importance of visual information, as well $1 - \rho$ for audio information. So we bring the transition factor φ_{tr} for audio transition to change the balance between ρ and $1 - \rho$:

$$\rho'(f) = \frac{\rho(f)}{\rho(f) + ((1 - \rho(f)) \cdot (1 + \varphi_{tr}(f))} = \frac{\rho(f)}{1 + \varphi_{tr}(f) - \rho(f) \cdot \varphi_{tr}(f)} \tag{15}$$

Because of φ_{tr} and $\tau(f_i, f_j)$, the fundamental formula of Balanced AV-MMR, Eq. 13, transforms into the following formula, which is defined as the formula of Balanced AV-MMR V1:

$$f_{k+1} = argmax_{f \in V \backslash S_k} \{\rho'(f) \cdot MR_I(f, S_k) + (1 - \rho'(f)) \cdot MR'_A(f, S_k)\} \tag{16}$$

3.3.3 Balanced AV-MMR V2 (using face detection)

According to the analysis in [25], the face is extremely important in visual information. Similar to Balanced AV-MMR V1, when the face appears in the video track of an audio segment, the video content becomes more important in the balance. Moreover, the face 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. Similar to Section 3.3.2 we linearly introduce this face factor to the balance and visual similarity as $\beta_{face} \cdot \rho$ and $\beta_{face} \cdot sim(f_i, f_j)$.

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, when the face appears in a video frame. After introducing the face factor β_{face} to $\rho'(f)$ in Section 3.3.2, it becomes:

$$\begin{aligned} \rho''(f) &= \frac{\rho(f) \cdot (1 + \beta_{face}(f))}{\rho(f) \cdot (1 + \beta_{face}(f)) + (1 - \rho(f)) \cdot (1 + \varphi_{tr}(f))} \\ &= \frac{\rho(f) \cdot (1 + \beta_{face}(f))}{1 + \varphi_{tr}(f) + \rho(f) \cdot (\beta_{face}(f) - \varphi_{tr}(f))} \end{aligned} \tag{17}$$

where $\beta_{face}(f) = facenumber(f) \cdot \theta_{face}$. 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. 12 evolve into:

$$\begin{aligned} &Sim'_{I1}(f_i, V \backslash S_k) = \tfrac{1}{|V \backslash (S_k \bigcup f_i)|} \textstyle\sum_{f_j \in V \backslash (S_k \bigcup f_i)} \beta'_{face}(f_i, f_j) sim(f_i, f_j); \\ &Sim'_{I2}(f, g) = \beta'_{face}(f, g) \cdot sim(f, g) \end{aligned} \tag{18}$$

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

Based on above development, Eq. 16 of Balanced AV-MMR V1 can be reformulated as:

$$f_{k+1} = argmax_{f \in V \backslash S_k} \{\rho''(f) \cdot MR'_I(f, S_k) + (1 - \rho''(f)) \cdot MR'_A(f, S_k)\} \tag{19}$$

where $MR'_I(f, S_k) = \lambda Sim'_{I1}(f, V \backslash S_k) - (1 - \lambda) \max_{g \in S_k} Sim'_{I2}(f, g)$.

3.3.4 Balanced AV-MMR V3 (adding temporal distance factor)

In a video two frames close temporally seem to be redundant. The frames in a video normally represent similar or relevant video content at the semantic level, while the frames from two different videos without the duplicate should represent less relevant content, even if they have the same similarity value according to low-level features. Therefore, 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 at the semantic level. It is possible that two frames far from each other in time represent similar visual content, especially in a video with multiple shots, but more frames around these two frames normally include more non-similar visual content, which could compensate the contrary influence caused by these two similar frames. Meanwhile, two neighbor frames may represent significantly different visual information in news video, but the audio information of these two frames, like the speech from the anchorman, is still highly relative at the semantic level. So these two frames in the news video are semantically very similar, which cannot be corrupted by the different visual information. Therefore, we argue that even in a set of multiple videos with multiple shots, our assumption still works.
- For multiple videos, a frame is more similar to another frame in the same video than a frame from another non-duplicated video. It is also possible that similar or same contents exist in different videos, but its contrary influence would be compensated by other normal frames.
- The factor of temporal distance in Balanced AV-MMR better distinguishes the summarization of multi-video from the summarization of multiple frames of multi-video. The temporal relation between videos are considered obviously different from the temporal relation between video frames in a 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

$$\alpha_{time}(f_i, f_j) = \begin{cases} 1, \text{if } f_i \text{ and } f_j \text{ are from two videos;} \\ 1 + \theta_{time} \cdot (1 - \frac{|t(f_i) - t(f_j)|}{10 * D_M}), \text{if } f_i \text{ and } f_j \text{ are from the same video.} \end{cases} \tag{20}$$

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. α_{time} is also a linear weight as τ and β_{face} and introduced in the way of $\alpha_{time}(f_i, f_j) \cdot sim(f_i, f_j)$. Then the similarities of the frames in Balanced AV-MMR become:

$$\begin{aligned} &Sim''_{I1}(f_i, V \backslash S_k) = \tfrac{1}{|V \backslash (S_k \bigcup f_i)|} \cdot \textstyle\sum_{f_j \in V \backslash (S_k \bigcup f_i)} \beta'_{face}(f_i, f_j) \alpha_{time}(f_i, f_j) sim(f_i, f_j); \\ &Sim''_{I2}(f, g) = \beta'_{face}(f, g) \alpha_{time}(f_i, f_j) sim(f, g); \\ &sim''_{A1}(f_i, A \backslash S_k) = \tfrac{1}{|A \backslash (S_k \bigcup f_i)|} \cdot \textstyle\sum_{f_j \in A \backslash (S_k \bigcup f_i)} \tau(f_i, f_j) \alpha_{time}(f_i, f_j) sim(f_i, f_j); \\ &sim''_{A2}(f, g) = \tau(f, g) \alpha_{time}(f_i, f_j) sim(f, g). \end{aligned} \tag{21}$$

Consequently, the formula of Balanced AV-MMR V3 is similar to Eq. 19 of Balanced AV-MMR V2 and generalized as

$$f_{k+1} = argmax_{f \in V \backslash S_k} \{\rho''(f) \cdot MR''_I(f, S_k) + (1 - \rho''(f)) \cdot MR''_A(f, S_k)\} \tag{22}$$

where

$$MR''_I(f, S_k) = \lambda Sim''_{I1}(f, V \backslash S_k) - (1 - \lambda) \max_{g \in S_k} Sim''_{I2}(f, g)$$

and

$$MR''_A(f, S_k) = \mu Sim''_{A1}(f, A \backslash S_k) - (1 - \mu) \max_{g \in S_k} Sim''_{A2}(f, g)$$

.

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 by [43] and the face by [35];
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

$$S_1 = argmax_{f_i}[\prod_{j=1,f_i \neq f_j}^{n} Sim_I(f_i, f_j) \cdot \prod_{j=1}^{n} Sim_A(f_i, f_j)]^{\frac{1}{n}}$$

   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 $S_{k+1} = S_k \bigcup \{f_{k+1}\}$;
6. Iterate to step 4 until S has reached the predefined size.

4 Experimental results

In this section, we will first describe the experimental video sets and the measures, Summary Reference Comparisons (SRCs) by Video Similarity (VS) - SRC_{VS} and SRC by Audio Video Similarity (AVS) - SRC_{AVS}, used to evaluate the summary. Then we introduce SRC to optimize the weight in MMR formulas. Later we compare Video-MMR summary to human summary, sparse representation summary, and K-means summary, and evaluate other MMR summaries globally for all the video genres and separately for different video genres. At last we discuss the experimental results in detail and suggest the best proposed algorithm for all the genres and for various genres.

4.1 Experimental videos and quality measures

We obtained our experimental video sets from a news aggregator website ”wikio.fr”. Totally we have 64 video sets plus a special video set, ”YSL”. A large scale corpus of 65 sets of videos, each set containing videos collected from various sources, but dealing with the same event. These video sets are classified into 7 genres: *Documentary, News, Music, Advertisement, Cartoon, Movie*, and *Sports*. Every set includes between 3 and 16 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. While, ”YSL” is composed of 14 videos, which are clustered according to the same topics, but whose video qualities, video genres and other properties of the videos inside them are various and diverse. The videos can be of the genre of movie, news, document, and others. The video inside it can be long duration like 8 minutes including lots of key scenes, or even the static image, while the total duration of ”YSL” is around 45 minutes. YSL videos can be downloaded and watched from: http://goo.gl/phpyDL. We have human summaries as ground truth for ”YSL” video set.

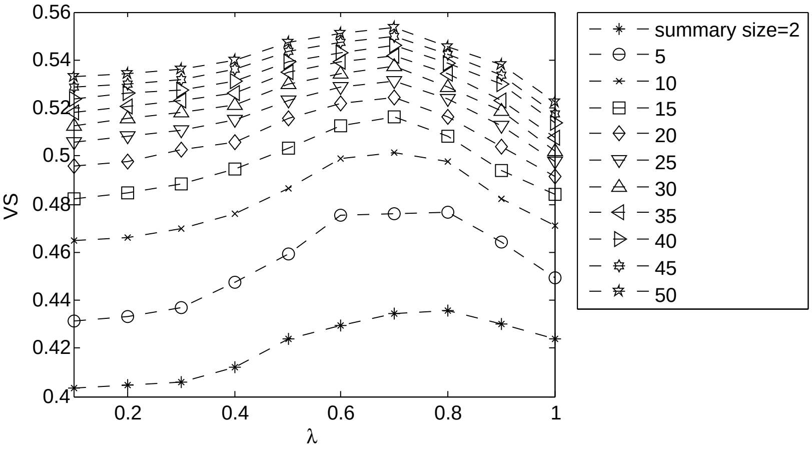

Fig. 2: SRC_{VS} for Video-MMR

To verify the effect of the proposed algorithms, we use SRC_{VS} and SRC_{AVS} between a summary and its original video sets to measure the quality of this summary. If SRC_{VS} or SRC_{AVS} of a summary is larger, the quality of this summary is better. SRC_{VS} is defined as

$$SRC_{VS}(S,V) = 1 - \frac{1}{n}\sum_{j=1}^{n} \min_{f_j \in V, g \in S}(1 - sim_I(f_j, g)) \tag{23}$$

And similarly SRC_{AVS} is defined as

$$SRC_{AVS}(S,V) = 1 - \frac{1}{n}\sum_{j=1}^{n} \min_{f_j \in V, g \in S}[1 - (sim_I(f_j, g) + sim_A(f_j, g))/2] \tag{24}$$

where n is the number of frames in V. g and f_j are frames respectively from video summary S and V. In the experiment, we select one frame per second. And we use Mel-frequency cepstral coefficients (MFCCs) feature to compute audio similarity and Bag of Word (BOW) for visual feature to compute visual similarity of the frames. The principle to get visual words is following: First we detect Local Interest Points (LIPs) in the frames, based on the Difference of Gaussian and Laplacian of Gaussian, to get a SIFT descriptor. Then SIFT descriptors are clustered into 500 groups by K-means to get the visual vocabulary with 500 words. Local Interest Point Extraction Toolkit [44] is exploited to quickly get BOW.

4.2 SRC

By sampling λ into 0.1, 0.2, 0.3, ..., 0.9, 1.0 in Video-MMR, we can implement SRC_{VS} of different weights. Fig. 2 shows SRC_{VS} of Video-MMR, whose summary sizes vary from 2 to 50 frames. SRC_{VS} is the average value of our experimental video sets. When $\lambda = 0.7$, SRC_{VS} is globally maximized. So in Video-MMR $\lambda = 0.7$. Similarly we optimize μ as 0.5 in AV-MMR, and Balanced AV-MMR variants.

4.3 The evaluation of video summaries

We compare Video-MMR with human summary as ground truth, the classic K-means and the state-of-the-art key frame extraction algorithm of sparse representation [20]. For K-means, we cluster the frames into the clusters by their features and select one frame from each cluster, which can represent its cluster, to form the summary. While for sparse representation, we use the same scenario as the K-means method. We manually test and choose the best parameters in the methods of K-means and sparse representation. We use the same features in all 3 approaches as described above in Section 4.1.

It is true that if we can use more human summary or assessment as the ground truth in the evaluation, the evaluation is more reliable. That is also what the researchers implement every year in the Trecvid [36][40]. However, because of our available time and collaborators, we have to limit our experiments but have designed an efficient way to implement the experiment. We select the video set, YSL, containing diverse kinds of videos with various durations as described above to do this experiment. And furthermore we choose 6 videos with the most obvious features and common contents, which can represent YSL [24] and are not too many to cause difficulty to the instant memory of the people. 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 by considering the factors of clearness and information coverage. 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 the weight, 0. Similar to Eq. 3, the summary quality of Video-MMR with respect to the human choice can be defined as:

$$QC_{Video\text{-}MMR} = \frac{1}{m}\sum_{i=1}^{m} w_i \cdot \max_{f \in S} sim(f, g_i) \tag{25}$$

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

$$QC_{human} = \frac{1}{N}\sum_{n=1}^{N} QC_n \tag{26}$$

where $QC_n = \frac{1}{m'}\sum_{i=1}^{m'} w_i \cdot \max_{f \in S_n} sim(f, g_i)$. In Eq. 25, $N = 12$, 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 qualities of Video-MMR, K-means, Sparse Representation, and human summaries (at least for a summary size of 10 keyframes). From Fig. 3, we can see that $QC_{Video\text{-}MMR}$ increases with the increase of summary size, because of more included information in the summary. Video-MMR is proved to be better than K-means, Sparse Representation and closer to ground truth. To further demonstrate the proposed Video-MMR, we display the video summaries with the size 10 of the video set "YSL" in Fig. 4. Each row is a summary from an algorithm. From top to bottom the summary of each row is from Video-MMR, K-means and Sparse Representation. Here the summary size is 10 frames, but the total duration of videos in "YSL" is much longer. As we found, though more frames can represent more video content, it is impossible for the viewers to remember and capture all the displayed visual information from these frames. Therefore, more frames do not mean that it is better for the viewers. After considering the tradeoff of representing more video content and the capability of the viewers, we use 10 frames as the smmary size to show it to the viewers. Also, it is impossible to represent all the visual information

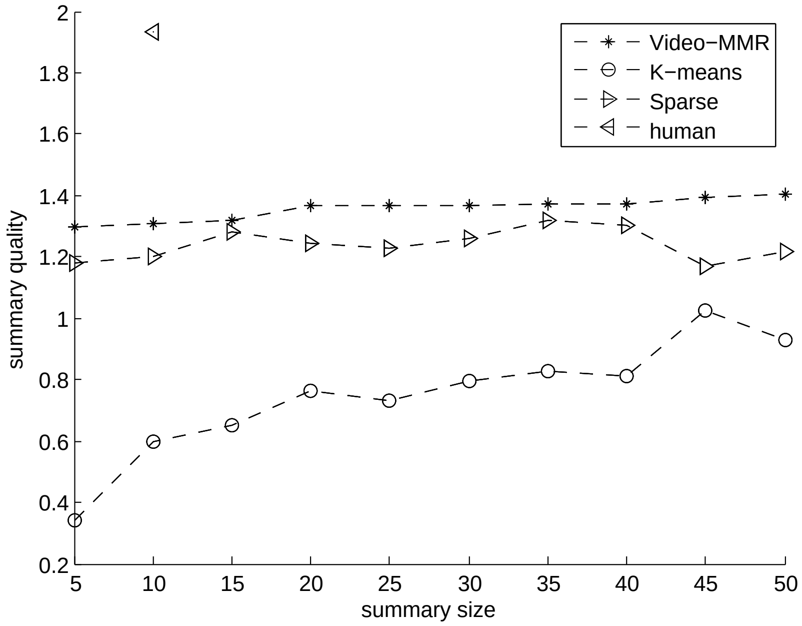

Fig. 3: Summary qualities of human, Video-MMR, Sparse Representation and K-means

in "YSL", but we hope that more important video contents are included in the summary with the size of 10. We can intuitively feel that, after watching the videos of "YSL", in the summaries with the size of 10, the summary by Video-MMR represents more visual information than Sparse Representation and even more than K-means. In the state-of-the-art subjective evaluation of video summary [18], people normally need to answer the quiz of clearness, information coverage and so on depending on the demand. In some papers [47], the authors need the people to consider all the quiz factors and only give a total rating score. In our case, we require people to implement the human assessment as the latter case. We request 7 people to assess each summary, and show rating scores and the mean rating score of 7 scores in Table 1. 10 is the highest and 0 is the lowest in the rating score. It is obvious that Video-MMR is rated the best by the human, better than K-means and Sparse Representation based algorithm in [20]. Though [10] is designed for the unstructured consumer videos and our proposed system is for the general videos, including the highly edited video set, it is still meaningful to compare them. The reason is that our video set in the experiment contains the videos with different genres, just like consumer videos, though the video set is classified into a genre. We also want to implement another algorithm based on sparse representation in [47]. But the scores of the face, image quality and so on are not carefully described in [47], so it is hard for us to exactly reimplement the experiment. However, in [47] the mean rating scores of the proposed algorithm is 17.4% higher than K-means, while in Table 1 the mean rating score of Video-MMR is more than 2 times K-means. So we can argue that for multi-video summarization Video-MMR is much more better than K-means and better than Sparse Representation based algorithm in [47] for multi-video summarization.

For the other Multimedia MMR algorithms, it is ideal to compare them to ground truth. But when considering audio information it is hard to get a coherent evaluation from a few people. And from Fig. 3

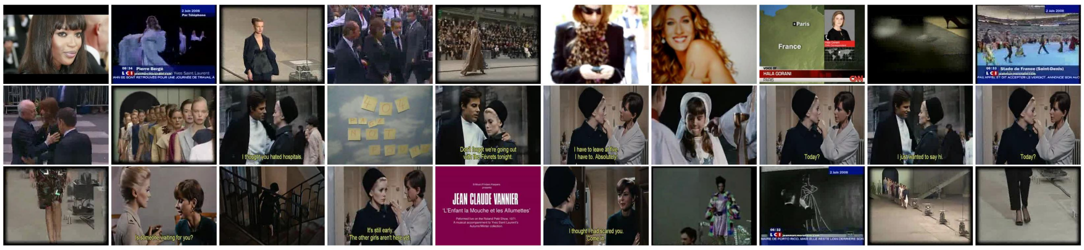

Fig. 4: The summaries with size 10. Top summary: Video-MMR; Middle summary: K-means; Bottom summary: Sparse Representation.

Table 1: Mean rating scores for ”YSL”

YSL	Video-MMR	K-means	Sparse Representation
Mean scores	8.286	3.714	5.857
Person 1	9	4	6
Person 2	8	5	7
Person 3	10	5	8
Person 4	7	1	3
Person 5	8	3	4
Person 6	8	5	8
Person 7	8	3	5

and Table 1, it is clear that the ranking of visual similarity and human assessment of different approaches are coherent. Therefore, we use SRCs of Video-MMR and the other MMR algorithms to demonstrate the other algorithms. If the similarity values of a proposed MMR algorithm in SRC is globally larger than Video-MMR for all the video summary sizes, we regard this algorithm better than Video-MMR and its summary closer to video set, and vice versa. Here we implement a large-scale analysis to 64 video sets in our experimental data of different video genres. We show SRCs of different algorithms globally for all video genres and separately for different genres in Fig. 5 by SRC_{VS} and Fig. 6 by SRC_{AVS}. When considering the effects of the algorithms using audio information, we should mainly consider SRC_{AVS} using both audio and visual similarity. We can find that in SRC_{AVS} figures of Fig. 6 the proposed algorithms, AV-MMR, and Balanced-MMR using audio information too are almost all better than Video-MMR.

4.4 Discussion

To summarize the video without audio track for all the genres, we don't have the choice to use the proposed algorithms by audio information too, so we have to use Video-MMR. But still in Fig. 5 we can find that Video-MMR is only a little worse than the best ones in SRC_{AVS} figures of different genres. And even for ”cartoon” and ”music” in Fig. 5, Video-MMR get the best values. So we can conclude that by Video-MMR we can get a good summary for the videos without audio information.

In ”movie” and ”sport” video sets of Fig. 5, Video-MMR performs worse than other algorithms, which is an abnormal phenomenon. The possible reason is that the factor of high dynamic motion is not considered in Video-MMR. While, other MMR algorithms with audio information compensate this limit because high dynamic motion often happens together with high pitch in the audio like the time of shoot

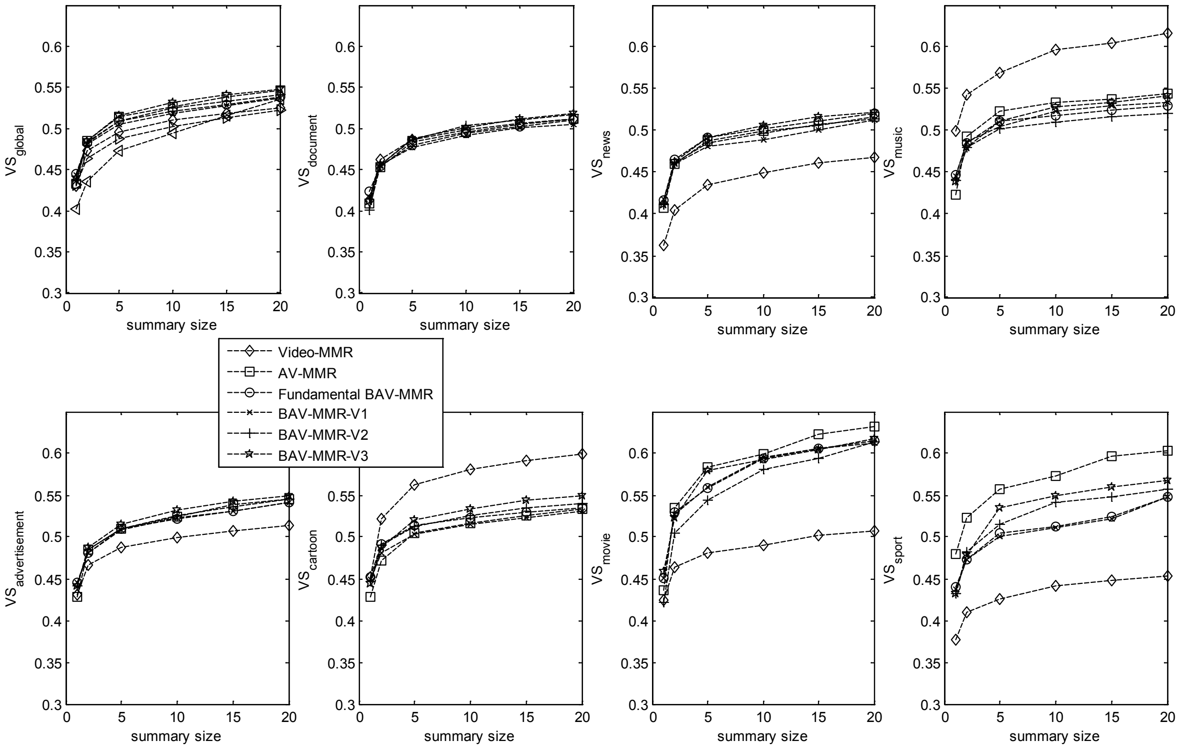

Fig. 5: SRC_{VS} by genre

in sports video. This also proves that more semantic-level feature can be used to improve the proposed algorithms of video summarization.

If we can utilize both visual and audio information in the video, Balanced AV-MMR V3 (BAV-MMR-V3) is the best in AVS_{global} for videos from all the genres. Consequently, BAV-MMR-V3 is suitable for the generic use when the genre of video set is not available. And for other figures in Fig. 6 BAV-MMR-V3 is also the best, which means that it is the optimal choice for all the genres and a global algorithm.

We can find that Video-MMR and AV-MMR significantly perform worse than other algorithms in Fig. 6. For Video-MMR it is easy to understand because Video-MMR is not considering audio information in the summarization but measured by SRC_{AVS}. The bad performance of AV-MMR justifies the difficulty to fuse the visual and audio information in video summarization mentioned in Section 1. The simple fuse of visual and audio information in AV-MMR achieves the worse video summaries than the variants of BAV-MMR. Therefore, the proposed fuse of video features in semantic level and feature level is a significant contribution to the community.

Finally, for the video and video sets of various genres we can use the algorithm of Video-MMR, when only the visual information is available, and BAV-MMR-3, when both visual and audio information is ready, to get the optimal video summaries compared to the state of the art.

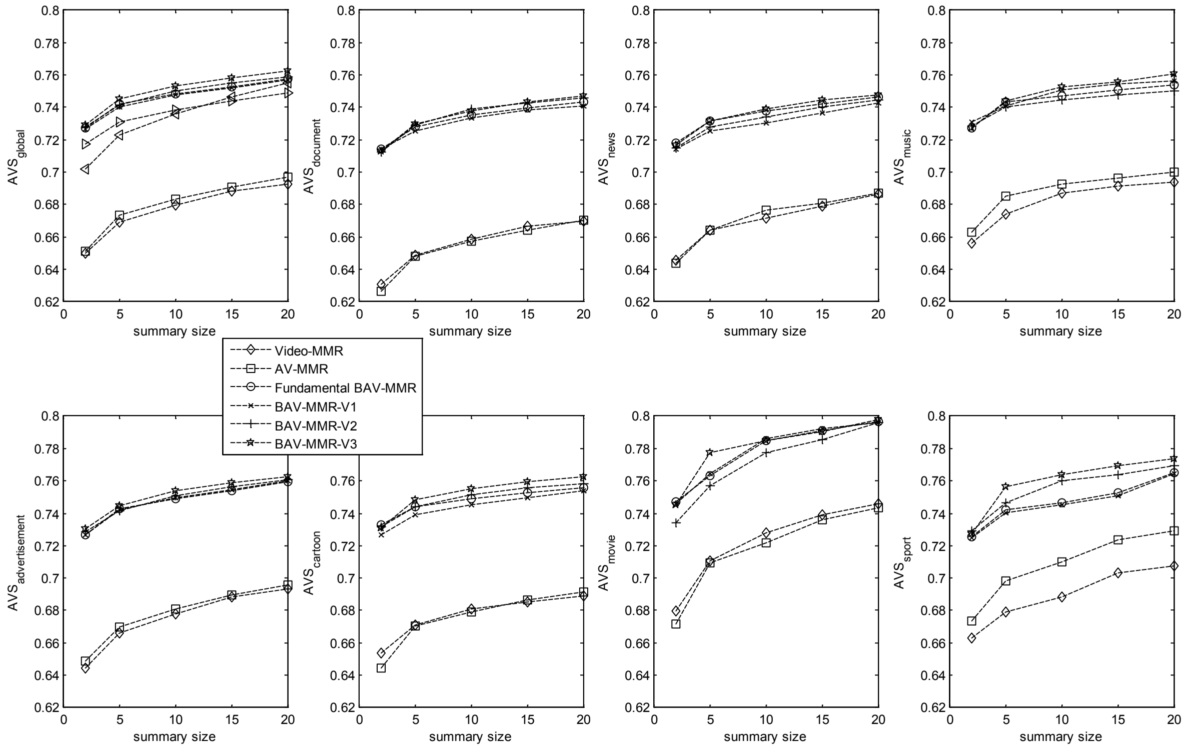

Fig. 6: SRC_{AVS} by genre

5 Conclusion

We have proposed a novel package of video summarization algorithms: Video-MMR, AV-MMR, and Balanced AV-MMR. Video-MMR borrows the idea from the successful MMR for text summarization. And Video-MMR is extended to other Multimedia MMR algorithms by considering both acoustic and visual information in the video. Through the experiments, we have proved Video-MMR by human assessment and the comparison to K-means and sparse representation based summarization. We also suggest the best video summarization algorithms: Video-MMR by only visual information, and Balanced AV-MMR by audio and visual information globally and for most genres.

References

1. Ajmal, M., Ashraf, M., Shakir, M., Abbas, Y., Shah, F.: Video summarization: Techniques and classification. Computer Vision and Graphics pp. 1–13 (2012)
2. Allen, M.J., Weintraub, L., Abrams, B.S.: Forensic Vision with Application to Highway Safety. Lawyers & Judges Publishing (2008)
3. Barbieri, M., Agnihotri, L., Dimitrova, N.: 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 (2003)
4. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of ACM SIGIR conference. Melbourne Australia (1998)

5. Chiu, P., Girgensohn, A., Polak, W., Rieffel, E., Wilcox, L.: A genetic algorithm for video segmentation and summarization. In: Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, vol. 3, pp. 1329–1332. IEEE (2000)
6. Cong, Y., Yuan, J., Luo, J.: Towards scalable summarization of consumer videos via sparse dictionary selection. Multimedia, IEEE Transactions on **14**(1), 66–75 (2012)
7. Dale, K., Shechtman, E., Avidan, S., Pfister, H.: Multi-video browsing and summarization. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pp. 1–8. IEEE (2012)
8. Das, D., Martins, A.F.: A survey on automatic text summarization. Tech. rep., Literature Survey for the Language and Statistics II course at CMU (2007)
9. Dimitrova, N.: Context and memory in multimedia content analysis. IEEE Multimedia 11 pp. 7–11 (2004)
10. Ding, D., Metze, F., Rawat, S., Schulam, P., Burger, S., Younessian, E., Bao, L., Christel, M., Hauptmann, A.: Beyond audio and video retrieval: towards multimedia summarization. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, p. 2. ACM (2012)
11. Dreyfus, H.L., Drey-fus, S.E., Zadeh, L.A.: Mind over machine: The power of human intuition and expertise in the era of the computer. IEEE Expert **2**(2), 110–111 (1987)
12. Dumont, E., Merialdo, B.: Automatic evaluation method for rushes summary content. In: Proceedings of International Workshop on Content-Based Multimedia Indexing, pp. 451–457. London, UK (2008)
13. Ejaz, N., Mehmood, I., Wook Baik, S.: Efficient visual attention based framework for extracting key frames from videos. Signal Processing: Image Communication (2012)
14. Fraternali, P., Martinenghi, D., Tagliasacchi, M.: Top-k bounded diversification. In: Proceedings of the 2012 international conference on Management of Data, pp. 421–432. ACM (2012)
15. Furini, M., Ghini, V.: An audio-video summarization scheme based on audio and video analysis. Consumer Communications and Networking Conference (2006)
16. Gao, S., Tsang, I., Chia, L.: Kernel sparse representation for image classification and face recognition. Computer Vision–ECCV 2010 pp. 1–14 (2010)
17. Haroz, S., Whitney, D.: How capacity limits of attention influence information visualization effectiveness. IEEE Trans. Vis. Comput. Graph. **18**(12), 2402–2410 (2012). URL `http://dblp.uni-trier.de/db/journals/tvcg/tvcg18.html#HarozW12`
18. He, L., Sanocki, E., Gupta, A., Grudin, J.: Auto-summarization of audio-video presentations. In: Proceedings of the seventh ACM international conference on Multimedia (Part 1), pp. 489–498. ACM (1999)
19. Jiang, W., Cotton, C., Loui, A.: Automatic consumer video summarization by audio and visual analysis. In: Multimedia and Expo (ICME), 2011 IEEE International Conference on, pp. 1–6. IEEE (2011)
20. Kumar, M., Loui, A.: Key frame extraction from consumer videos using sparse representation. In: Image Processing (ICIP), 2011 18th IEEE International Conference on, pp. 2437–2440. IEEE (2011)
21. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. Advances in neural information processing systems **19**, 801 (2007)
22. Lew, M., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Transactions on Multimedia Computing (2006)
23. Li, Y., Merialdo, B.: Multi-video summarization based on Video-MMR. In: Proceedings of 11th International Workshop on Image Analysis for Multimedia Interactive Services. Desenzano del Garda, Italy (2010)
24. Li, Y., Merialdo, B.: VERT: automatic evaluation of video summaries. In: Proceedings of ACM Multimedia Conference. Firenze, Italy (2010)
25. Li, Y., Merialdo, B.: Multi-video summarization based on Balanced AV-MMR. In: Proceedings of The 18th International Conference on MultiMedia Modeling. Klagenfurt, Austria (2012)
26. Lienhart, R., Pfeiffer, S., Effelsberg, W.: Video abstracting. Communications of the ACM 40 (12) pp. 55–62 (1997)
27. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). Barcelona, Spain (2004)
28. Lin, K., Lee, A., Yang, Y., Lee, C., Chen, H.: Automatic highlights extraction for drama video using music emotion and human face features. In: Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on, pp. 1–6. IEEE (2011)
29. Ma, Y., Hua, X., Lu, L., Zhang, H.: A generic framework of user attention model and its application in video summarization. IEEE Transactions on Multimedia 7 pp. 907–919 (2005)
30. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia, pp. 533–542. ACM (2002)
31. Marois, R., Ivanoff, J.: Capacity limits of information processing in the brain. Trends in Cognitive Sciences **9**(6), 296–305 (2005)
32. McDonald, R.: A study of global inference algorithms in multi-document summarization. Advances in Information Retrieval pp. 557–564 (2007)

33. Mckeown, K., J.Passonneau, R., K.Elson, D.: Do summaries help? a task-based evaluation of multi-document summarization. In: Proceedings of ACM SIGIR conference. Melbourne Australia (1998)
34. Money, A.G., Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art. Journal of visual communication and image representation (2007)
35. Nilsson, M., Nordberg, J., Claesson, I.: Face detection using local smqt features and split up snow classifier. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (2007)
36. Over, P., Smeaton, A.F., Kelly, P.: The trecvid 2007 bbc rushes summarization evaluation pilot. In: Proceedings of ACM MM'07. Augsburg, Bavaria, Germany (2007)
37. Peng, W., Chu, W., Chang, C., Chou, C., Huang, W., Chang, W., Hung, Y.: Editing by viewing: automatic home video summarization by viewing behavior analysis. Multimedia, IEEE Transactions on **13**(3), 539–550 (2011)
38. Rudinac, S., Larson, M., Hanjalic, A.: Learning crowdsourced user preferences for visual summarization of image collections (2013)
39. Shapiro, K.E.: The limits of attention: Temporal constraints in human information processing. Oxford University Press (2001)
40. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM Press, New York, NY, USA (2006). DOI http://doi.acm.org/10.1145/1178677.1178722
41. Sugano, M., Nakajima, Y., Yanagihara, H.: Automated MPEG audio-video summarization and description. In: Proceedings of The International Conference on Image Processing. New York, USA (2002)
42. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Transaction Multimedia Compuatation Communication Application 3 (2007)
43. University of Cambridge: HTK toolkit. `http://htk.eng.cam.ac.uk`
44. Video Retrieval Group, City U. of Hong Kong: Local interest point extraction toolkit. `http://vireo.cs.cityu.edu.hk`
45. Wactlar, H.D.: Multi-document summarization and visualization in the informedia digital video library. In: Proc. of the 12th New Information Technology Conference. Beijing, China (2001)
46. Wang, F., Merialdo, B.: Multi-document video summarization. In: Proceedings of International Conference on Multimedia and Expo. New York, USA (2009)
47. Wang, Z., Kumar, M., Luo, J., Li, B.: Sequence-kernel based sparse representation for amateur video summarization. In: Proceedings of the 2011 joint ACM workshop on Modeling and representing events, pp. 31–36. ACM (2011)
48. Xu, C., Shao, X., Maddags, N.C., Kankanhalli, M.S.: Automatic music video summarization based on audio-visual-text analysis and alignment. ACM SIGIR (2005)
49. Yahiaoui, I., Merialdo, B., Huet, B.: Automatic video summarization. Multimedia Content-based Indexing and Retrieval (2001)
50. Yang, C.C., Chen, H., Hong, K.: Visualization of large category map for internet browsing. Decision Support Systems **35**(1), 89–102 (2003)