Multi-Video Summarization Based on OB-MMR

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

Abstract

In this paper we propose a novel algorithm for video summarization, OB-MMR (Optimized Balanced Audio Video Maximal Marginal Relevance). This algorithm is suitable to summarize both single and multiple videos. OB-MMR is achieved by optimizing the parameters in Balanced AV-MMR (Balanced Audio Video Maximal Marginal Relevance), namely the balance factor between audio information and visual information in the video, but also the importance of face and audio transitions among audio segments with different genres. Therefore, OB-MMR achieves a better result than previous algorithms, Video-MMR and Balanced AV-MMR. Furthermore, it is possible to select the optimized parameters for each genre of videos, which leads to promising automatic algorithms for video summarization in the future large-scale experiments.

1. Introduction

With the spectacular increase of videos from TV, mobile phone, and Internet, human beings cannot handle the explosion of multimedia information. It is necessary for a person to spend a lot of time to find his interesting videos. Therefore, various methods, like shot detection and video abstraction, have been proposed by researchers to facilitate the access to large quantities of video information. Among these methods, video summarization has an important role. Video summarization produces summaries by analyzing the video content, and condenses this content into an abbreviated descriptive form. Video summaries can be for example used in interactive browsing and searching systems, by which the user easily manages and accesses the digital video content.

Earlier work in video summarization focused on processing a single video [1] [2], while more recent approaches have considered the case of multi-video summarization [3] [4]. However there are still many limitations. Many existing algorithms only consider the features from the video track, and neglect the audio track because of the difficulty of combining the information from audio and video. Some algorithms like [5] consider both the audio track and video track, but they are often domain-specific. There are not many generic algorithms, because in this case some specific features like music energy are difficult to use in a generic manner.

In previous work, the authors have proposed a series of generic algorithms, Video-MMR (Video Maximal Marginal Relevance) [4] for multi-video summarization by using only visual information, AV-MMR (Audio Video Maximal Marginal Relevance) by exploiting both audio and visual information [10], and Balanced AV-MMR (Balanced Audio Video Maximal Marginal Relevance) [11] which considers the balance factor between audio and visual information. In this paper, we improve over Balanced AV-MMR by optimizing some parameters in the algorithm.

This paper is organized as follows: Section 2 briefly reviews video summarization. Then Section 3 introduces related work of MMR. And Section 4 and 5 describe the principle and experimental results of OB-MMR. At last the paper is concluded in Section 6.

2. The review of video summarization

The aim of video summarization is to exploit the audiovisual information to obtain the underlying relation among video frames and create a condensed version from the original videos. Video summaries are useful especially for long videos, which could save the time and help the user understand the whole video without watching it honestly in details. The user could choose the most interesting videos faster and easier. Another possible advantage of video summaries is to prepare the original video for the searching engine. Only the selected frames of videos are inserted in the index, which saves a great amount of resources. The methods for video summarization are various, but some basic features are very often present. Four audiovisual cues may be used in video summaries [2]: keyframes, text, video segments and graphic cues. In video summarization, it is still a problem to perfectly combine the information from text, audio track and video track, though some algorithms [5] [11] [17] have been proposed.

The computational mechanism in video summarization [16] includes maximum content coverage, minimum correlation among summary elements, and interesting/highlight events. They are the aims for video summarization to achieve, no matter what kind of algorithm it is.

Video summaries can be distinguished by content types into object based, event based, perception and/or feature based summaries [2]. 1) Object based summaries focus on specific objects that occur within the video, especially human face, like [11]. In [11], the frames with faces own greater weights in the computation. 2) Event based summaries aim to find specific events in the video. Attention indicator and saliency map are significant measures in [19]. 3) Perceptions based summaries focus on high-level concepts and try to mimic the perception of the video by the user. For example, [18] summarizes the video by maximizing the entropy among concept entities. 4) Feature summarization tries to objectively generate the summaries from an analysis of low-level features, such as speech, color, texture, and the most duplicated scenes. In [17] the authors generate the skimming by intercepting the sections with the high sum of different feature curves.

There are usually two kinds of representations forms for video summaries: static video keyframes and dynamic video skimming [16] [17]. From these two forms, it is necessary to consider the best methods to display video summaries, such as video panel [19], story board, circle representation and graph tree [18].

3. Related work

3.1. MMR in Text Summarization

In the domain of Natural Language Processing, Maximal Marginal Relevance (MMR) proposed by J. Carbonell and J. Goldstein [9] is a successful algorithm for text summarization. MMR is based on the idea of Marginal Relevance (MR). In the case of the selection of documents while answering a query, the MR of a document with respect to the query Q and the current selection S is defined by:

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

Where Q is a query or user profile, and D_i and D_j are text documents in the returned list of documents R for the query Q. D_j is a document already selected in S, while D_i is a candidate in the list of unselected documents $R\setminus S$. By iteratively selecting the text fragments with MMR in the text document, a text summary can easily be constructed by Eq. 2:

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

3.2. Video-MMR

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

$$\begin{aligned} &\textit{Video-MR}(f) \\ &= \lambda\, Sim_1(f, V\setminus S) - (1-\lambda) \max_{g \in S} Sim_2(f, g) \quad (3) \end{aligned}$$

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

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

3.3. AV-MMR

Video-MMR is extended to AV-MMR in [10] by considering information from both audio and video. We then extend Eq. 4 to Eq. 5, which defines how summary S_{k+1} can be constructed by iteratively selecting a new audio-video segment:

$$\begin{aligned} &S_{k+1} = S_k \cup \underset{f \in V\setminus S_k}{arg\ max}[\\ &\lambda\, Sim_{I1}(f, V\setminus S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f, g) + \\ &\mu\, Sim_{A1}(f, V\setminus S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f, g)] \quad (5) \end{aligned}$$

where visual similarities Sim_{I1} and Sim_{I2} are the same measures as Sim_1 and Sim_2 in Eq. 4. Audio similarities Sim_{A1} and Sim_{A2} play roles similar to Sim_{I1} and Sim_{I2}. Eq. 5 combines visual and audio similarities corresponding to the same frame, and it is called Synchronous AV-MMR.

3.4. Balanced AV-MMR

In [11], the authors consider that audio is composed by audio segments corresponding to silence, music, and speech. The HTK toolkit [7] is used to automatically detect those genres of audio segments. Several remarks are used to better combine audio and video information. When the audio information is significantly changing, it is likely that the user will pay more attention to this instant. The importance of video track and audio track in an audio segment are complementary, so a combination of the summary created by video track and the summary created by audio track is utilized in Balanced AV-MMR.

3.4.1. Fundamental Balanced AV-MMR. The fundamental formula of Balanced AV-MMR is:

$$f_{k+1} = arg\ max_{f \in V \backslash S_k} \{\rho_T(f) \\ [\lambda\, Sim_{I1}(f, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g)] \\ + \left(1 - \rho_T(f)\right) \\ [\mu\, Sim_{A1}(f, V \backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{A2}(f,g)]\} \quad (6)$$

We define the importance ratio ρ between audio summary and video summary as: $\rho_T(f) = \frac{N_V(f)}{N_V(f)+N_A(f)}$, where $N_V(f)$ and $N_A(f)$ are summary sizes of video summary, and audio summary in audio segment T where frame f is inside.

3.4.2. Balanced AV-MMR V1. Then we introduce the augment factor τ for audio genre. The importance ratio becomes from ρ to ρ' : $\rho'_T(f) = \frac{\rho_T(f)}{\rho_T(f)+((1-\rho_T(f))+\varphi_{tr})}$, where φ_{tr} is the factor brought by the transition of audio genres. And the formula of BAV-MMR V1 is modified to the following formula:

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

3.4.3. Balanced AV-MMR V2. After introducing face information β_{face} to Eq. 7, the formula of Balanced AV-MMR V2 is:

$$f_{k+1} = arg\ max_{f \in V \backslash S_k} \{\rho''_T(f) \\ [\lambda\, Sim'_{I1}(f, V \backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim'_{I1}(f,g)] \\ +(1 - \rho''_T(f)) \\ [\mu\, Sim'_{A1}(f, V \backslash S_k) - (1-\mu) \max_{g \in S_k} Sim'_{A2}(f,g)]\} \quad (8)$$

where $\rho''_T(f) = \frac{\rho_T(f)+\beta_{face}(f)}{(\rho_T(f)+\beta_{face}(f))+((1-\rho_T(f))+\varphi_{tr})}$.

3.4.4. Balanced AV-MMR V3. At last, we consider that it is necessary to consider the temporal distance when computing the similarity between two frames f_i and f_j. Consequently, the formula of Balanced AV-MMR V3 is the same with Eq. 8 of Balanced AV-MMR V2, but Sim'_{I1} and Sim'_{A1} contain the factor of temporal distance α_{time}: $Sim'_{I1}(f_i, V \backslash S_k) = \frac{1}{|V \backslash (S_k \cup f_i)|} \cdot \sum_{f_j \in V \backslash (S_k \cup f_i)} \beta_{face}(f_i, f_j) \alpha_{time}(f_i, f_j) sim(f_i, f_j)$ and $Sim'_{A1}(f_i, V \backslash S_k) = \frac{1}{|V \backslash (S_k \cup f_i)|} \sum_{f_j \in V \backslash (S_k \cup f_i)} \tau(f_i, f_j) \alpha_{time}(f_i, f_j) sim(f_i, f_j)$. Sim'_{I2} and Sim'_{A2} own the factor α_{time} similarly, and

$$\alpha_{time}(f_i, f_j) = \begin{cases} 1.1 & , if\ f_i\ and\ f_j\ are\ from\ two\ videos; \\ 1 + \frac{|t(f_i)-t(f_j)|}{10 * D_M} & , if\ two\ are\ from\ the\ same\ video \end{cases}.$$

$t(f_i)$ and $t(f_j)$ are the frame times of f_i and f_j in video M. $facenumber_f$ means the number of faces in frame f. D_M is the duration of video M.

4. OB-MMR

In [11], Balanced AV-MMR exploits many parameters which are manually set according to experience. These parameters are: the balance parameter between audio and visual information ρ''_T, the parameter of temporal distance α_{time}, the face parameter β_{face}, the audio genre parameter τ, and the parameter for audio transition φ_{tr}.

However, it is hard to manually decide the best values of these 5 parameters. Also for different genres of videos, the optimal values of those parameters may vary, because the relation between video track and audio track is different. Therefore, we wish to propose an automatic mechanism to optimize the set of weights for Balanced AV-MMR. First, what we do is to reformulate Eq. 8 into the following formula:

$$f_{k+1} = arg\ max_{f \in V \backslash S_k} \left\{ \begin{array}{c} \rho''_T(f) \begin{bmatrix} \lambda\, Sim_{I1}{}'(f, V \backslash S_k) - \\ (1-\lambda) \max\limits_{g \in S_k} Sim_{I2}{}'(f,g) \end{bmatrix} + \\ (1 - \rho''_T(f)) \begin{bmatrix} \mu Sim_{A1}{}'(f_i, V \backslash S_k) - \\ (1-\mu) \max\limits_{g \in S_k} Sim_{A2}{}'(f,g) \end{bmatrix} \end{array} \right\} \quad (9)$$

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

$Sim_{I1}{}'(f_i, V \backslash S_k) = \frac{1}{|V \backslash (S_k \cup f_i)|} \sum_{f_j \in V \backslash (S_k \cup f_i)} [1 + (F(f_i) + F(f_j)) \cdot W_f](1 + T(f_i, f_j) \cdot W_t) sim(f_i, f_j)$;

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

$Sim_{A1}{}'(f_i, V \backslash S_k) = \frac{1}{|V \backslash (S_k \cup f_i)|} \sum_{f_j \in V \backslash (S_k \cup f_i)} (1 + S(f_i, f_j) \cdot W_s)(1 + T(f_i, f_j) \cdot W_t) sim(f_i, f_j)$;

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

Inside Eq. 9, the functions $B, F, T, R,$ and S are the computed features for the balance between audio and visual information, the face importance, the temporal distance, the audio transition, and the audio genre; while, W_b, W_f, W_t, W_r, and W_s are the weights for those features B, F, T, R, and S. Compared to Eq. 8, Eq. 9 is easier to be optimized because we just have to automatically adjust the values of the weights W_b, W_f, W_t, W_r, and W_s to achieve the best result. The result of the optimization of Eq. 9 is called Optimized Balanced AV-MMR (OB-MMR).

Before adjusting the weights in Eq. 9, we need to define the fitness function for these weights. In video summarization, we usually regard the summaries from human being as the ground truths, because video summarization is a problem which is absolutely human oriented. Assume that we already have some groups of human summaries, we could use the similarities between the summaries from OB-MMR and the summaries from human as the fitness function to adjust the weights W_b, W_f, W_t, W_r, and W_s as the fitness function, because we want the summary from OB-MMR more similar to the summary from human.

Then it is necessary to select an automatic algorithm to automatically tune the weights W_b, W_f, W_t, W_r, and W_s. One successful algorithm is Particle Swarm Optimization (PSO) proposed by R. Poli, J. Kennedy and T. Blackwell [12] [13]. PSO has been used across a wide range of applications, which has proved the effect of PSO. In PSO, every particle decides its movement by considering its current location and the previous best location of the particles. The individual contains three *D*-dimensional vectors: the current position $\vec{x}_i$, the previous best position $\vec{p}_i$, and the velocity $\vec{v}_i$. The best function result is denoted by $pbest_i$, and $\vec{p}_g$ is the best neighbor of $\vec{p}_i$. PSO procedure is described in [13]. In OB-MMR, 5 weights W_b, W_f, W_t, W_r, and W_s can be considered as 5 elements of a vector $\vec{x}_i$ in the searching space of PSO for the fitness function.

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

The optimized weights from PSO and GR will be shown in the next section. After we obtain the optimized weights, we could implement OB-MMR process in a way similar to Balanced AV-MMR:

Algorithm: OB-MMR

```
1: Summarize video track by Video-MMR with Eq. 4.
2: Summarize the audio track by Audio-MMR:
   $S_{k+1} = S_k \cup \underset{f \in A \backslash S_k}{argmax} \left(1 + S(f,g) \cdot W_s\right) \cdot \left(\begin{matrix} \lambda\, Sim_1(f, A \backslash S_k) - \\ (1-\lambda) \max\limits_{g \in S_k} Sim_2(f,g) \end{matrix}\right)$
3: Detect the audio segments and their genres by HTK audio system.
4: The initial video summary $S_1$ is initialized with one frame, defined as:
   $S_1 = arg \underset{f_i, f_i \neq f_j}{max} [\prod_{j=1}^{n} Sim_I(f_i, f_j) \prod_{j=1}^{n} Sim_A(f_i, f_j)]^{\frac{1}{n}}$
   where $f_i$ and $f_j$ are frames in video set $V$, and $n$ is the total number of frames except $f_i$. $Sim_I$ computes similarity of image information between $f_i$ and $f_j$, while $Sim_A$ is the similarity of audio information between $f_i$ and $f_j$.
5: Find the optimized weights for Eq. 9 by fitting the OB-MMR summary to human summaries.
6: loop
7:     Select frame $f_{k+1}$ by Eq. 9 with optimized weights
8:     Set $S_{k+1} = S_k \cup \{f_{k+1}\}$.
9.     Iterate to step 7) until $S$ has reached the predefined size.
10: End loop
```

5. Experimental results

We have two video sets, “DATI” and “YSL”, which are both news videos obtained from the aggregation website, “WIKIO”. There are 16 videos in “DATI”, and 14 videos in “YSL”. Both video sets own videos with the duration from around 30 seconds to around 10 minutes, and video categories vary from news, advertisements and music video to movie.

The visual content of a keyframe is represented by the Bag-Of-Word feature vectors [4] [15], and audio feature are MFCC vectors obtained by [6]. The similarities between audio features and visual features are identical to their definition in AV-MMR. And visual similarity is used as the fitness function in OB-MMR.

Same with Balanced AV-MMR, we also use the trained HTK toolkit [7] to process the audio track and get the audio genre of each audio frame. The statistical data of audio genres of audio frames in “DATI” and “YSL” is shown in Table 1. The toolkit provided in [8] is used to detect faces in the video frames. Furthermore, in [14] we have already got the human summaries for "DATI" and "YSL", which are used in the fitness function of PSO and GR to get the optimized weights for OB-MMR. The human summaries used as the ground truth are assessed by 12 people with professional background of image processing. Each person selects 10 most important video frames for each video in our prepared frames.

Table 1. The number of audio frames with different genres

	silence	music	speech
DATI	24	524	2366
YSL	57	1173	1318

In PSO, we consider a population, individual sets of weights, of size 20. And according to [13] $w = 0.7298$, and $\emptyset_1 = \emptyset_2 = 1.49618$ in Eq. 9. In the gridding, the range of gridding for each weight is $[0.0, 2.0]$ and the initial interval of gridding is 0.5. During the relaxation phrases, the intervals of the two relaxation iterations are 0.04 and 0.006 respectively.

And in Eq. 9 of OB-MMR, we define the parameters T, F, B, R and S as follows:

$$T(f_i, f_j) = \begin{cases} 1.1 & , if\ f_i\ and\ f_j\ from\ two\ videos \\ 1 + 0.1 \cdot \frac{|t(f_i) - t(f_i)|}{T_M} & , if\ f_i\ and\ f_j\ from\ same\ video \end{cases}$$

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

$F(f) = face\ number\ in\ frame\ f$;

$B(f) = \frac{N_V(f)}{N_V(f) + N_A(f)}$, where $N_V(f)$ is frame number of video summary, created by Video-MMR, in audio segment L where frame f is inside, and $N_A(f)$ is frame number of audio summary, by Audio-MMR, in L;

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

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

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

Table 2. Optimized weights and similarities

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

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

Table 3. Cross validation

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

From Table 3, it is obvious that the weights from PSO and GR for "DATI" are better than the weights from PSO and GR for "YSL" for both video sets, while the similarities are very similar in Table 2. Therefore we exploit these two sets of optimized weights, from PSO and GR, of "DATI" other than "YSL" in the following experiments.

Then we use these 2 weights to compute video distance, $d_{VD}(S, V) = \frac{1}{n} \sum_{j=1}^{n} \min_{f_j \in V, g \in S} (1 - sim'_I(f_j, g))$, and audio video distance $d_{AVD}(S, V) = \frac{1}{n} \sum_{j=1}^{n} \min_{f_j \in V, g \in S} \left[1 - \frac{sim'_I(f_j, g) + sim'_A(f_j, g)}{2}\right]$ between OB-MMR summaries and original videos [11], which are shown in Figure 1 and Figure 2. It is obvious that two OB-MMR curves are better than the previous algorithms, Video-MMR and Balanced AV-MMR. OB-MMR by PSO is a little better than OB-MMR by GR, which is caused by the better similarity with the human summaries shown in Table 2. And even when the optimized weights from video set "DATI" are used in OB-MMR for "YSL", the results in Figure 1 and Figure 2 are better. So OB-MMR is a generic algorithm, even the optimized weights are from the other videos. In the future, it is possible to decide a fixed optimized set of weights for each genre of videos after the large-scale experiments.

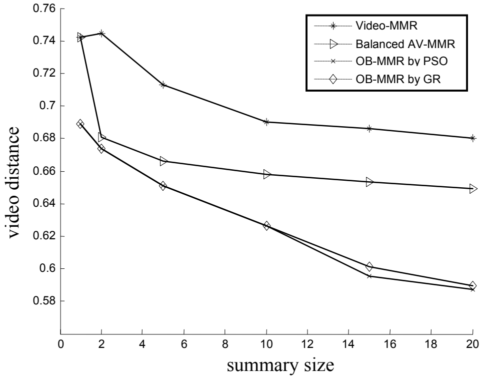

Figure 1. Video distance with original videos

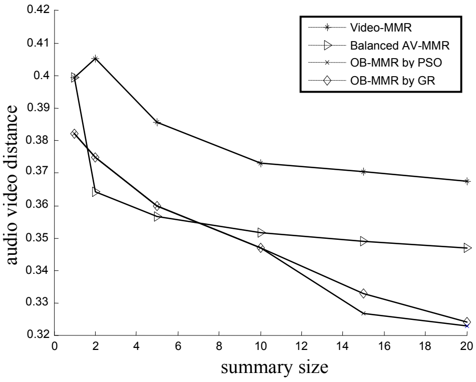

Figure 2. Audio video distance with original videos

We could also conclude that PSO is better than GR for OB-MMR according to the curves shown in Figure 1 and Figure 2. Furthermore, it is unnecessary for PSO to define the range and intervals to do the gridding and relaxation before the computation, so PSO is an unsupervised algorithm and better for OB-MMR. OB-MMR optimized by PSO can be used to summarize different categories of videos without a prior knowledge except the video category.

6. Conclusions

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

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

7. 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*, 121-143, 2008.

[3] F. Wang and B. Merialdo, "Multi-document Video Summarization", *International Conference on Multimedia & Expo*, 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*, Desenzano del Garda, Italy, 2010.

[5] Marco Furini and Vittorio Ghini, "An Audio-Video Summarization Scheme Based on Audio and Video Analysis", *IEEE Consumer Communications and Networking Conference*, USA, 2006.

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

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

[8] 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.

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

[10] Y. Li, B. Merialdo, "Multi-video Summarization Based on AV-MMR", *International Workshop on Content-based Multimedia Indexing*, France, 2010.

[11] Yingbo Li, Bernard Merialdo, "Video Summarization Based on Balanced AV-MMR", *Submitted to ICME VCIDS workshop*, Spain, 2011.

[12] James Kennedy, Russell Eberhart, "Particle swarm optimization", *In Proceedings of the IEEE international conference on neural networks IV*, pp. 1942–1948, 1995.

[13] R. Poli, J. Kennedy, Tim Blackwell, "Particle swarm optimization: An overview", *Swarm Intell*, 1: 33–57, 2007.

[14] Yingbo Li, Bernard Merialdo, "VERT: automatic evaluation of video summaries", *ACM Multimedia Conference*, Florence, Italy, 2010.

[15] http://vireo.cs.cityu.edu.hk.

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

[17] Y. Ma, L. Lu, H. Zhang and M. Li, "A User Attention Model for Video Summarization", *ACM International Conference on Multimedia*, USA, 2002.

[18] B. Chen, J. Wang, and J. Wang, "A Novel Video Summarization Based on Mining the Story-Structure and Semantic Relations Among Concept Entities", *IEEE Transactions on Multimedia*, VOL. 11, NO. 2, February 2009.

[19] C. Ngo, Y. Ma, H. Zhang, "Video Summarization and Scene Detection by Graph Modeling", *IEEE Transactions on Circuits and Systems for Video Technology*, VOL. 15, NO. 2, February 2005.