MULTI-VIDEO SUMMARIZATION BASED ON VIDEO-MMR

Yingbo Li and Bernard Merialdo

Institut Eurecom, France
{Yingbo.Li, Bernard.Merialdo}@eurecom.fr

ABSTRACT

This paper presents a novel and effective approach for multi-video summarization: Video Maximal Marginal Relevance (Video-MMR), which extends a classical algorithm of text summarization, Maximal Marginal Relevance. Video-MMR rewards relevant keyframes and penalizes redundant keyframes, as MMR does with text fragments. Two variants of Video-MMR are suggested, and we propose a criterion to select the best combination of parameters for Video-MMR. Then, we compare two summarization strategies: Global Summarization, which summarizes all the individual videos at the same time, and Individual Summarization, which summarizes each individual video independently and concatenates the results. Finally, Video-MMR algorithm is compared with popular K-means algorithm, supported by user-made summary.

1. INTRODUCTION

The video data on Internet and home equipments is growing day by day. In recent years, the number of Internet videos is unimaginably increasing with the development of online video websites, such as YouTube. Every day many people upload and share news videos, personal videos and so on. How to manage such a large amount of visual data is a serious problem for human beings, so it is an active research topic nowadays. Video summarization has been identified as an important component to deal with video data. Video summarization can select relevant keyframes or segments in videos and create a video summary, which contains the essential content of the video. While a lot of effort has been devoted to the summarization of a single video [2], less attention has been given to the summarization of a set of videos [1]. With the increase in quantity, it is more and more often that videos are organized into groups, for example on news websites, therefore the issue of creating a summary for a set of videos is getting an increased importance. This follows the trend that is now well established in the text document community, where multi-document summarization has been extensively studied [3] [4] [5] [9].

This paper proposes a novel summarization algorithm, Video-MMR extending the current text summarization algorithm MMR to video domain. We also explore the issue of Global Summarization (all videos together) versus Individual Summarization (each video one by one) in the construction of the summary. This paper is organized as follows: Section 2 presents the theory of Video-MMR in detail, together with original MMR of text summarization, and proposes a comparison method with user-made summaries as ground truth. In section 3, we use a criterion of the minimum distance with the original video to experimentally select the best variant of Video-MMR; we compare Global and Individual Summarization, and conclude that Global Summarization is better; meanwhile, summary quality of Video-MMR is assessed with human-generated ground truth, and compared with K-means algorithm. At last, we conclude this paper with some final remarks.

2. VIDEO MAXIMAL MARGINAL RELEVANCE

2.1. Text summarization and MMR

Text summarization is a hot research topic in the area of Natural Language Processing [5] [7] [8]. Text summaries preserve important information, and are short compared with original single document or multiple documents. Since 1990s, a lot of work is dedicated into the research of text summarization for multiple documents [5] [10]. Various approaches have been proposed, such as information fusion, graph spreading activation, centroid based summarization and multilingual multi-document summarization. A popular and efficient one is Maximal Marginal Relevance (MMR) proposed by J. Carbonell and J. Goldstein [6]. The Marginal Relevance (MR) of a document with respect to a query Q and a selection S is defined by the equation:

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

Where Q is a query or user profile, and D_i and D_j are text documents in 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. Marginal Relevance can be used to construct multi-document summaries by considering the set of all documents as the query Q, R as a set of text fragments, and iteratively selecting the text fragment D_{MMR} that maximizes the MR with the current summary:

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

2.2. Video summarization

The goal of Video summarization is to identify a small number of keyframes or video segments which contain as much information as possible of the original video. Video segments can be characterized by one or several keyframes, so we focus here on the selection of relevant keyframes. The relation between the visual content in the summary, *S*, and in the original video, *V*, can be measured by the following distance:

$$d(S,V) = \frac{1}{n} \sum_{j=1}^{n} \min_{f_j \in V, g \in S} [1 - sim(f_j, g)] \qquad (3)$$

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

$$\hat{S} = \underset{S}{argmin}[d(S,V)] \qquad (4)$$

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

2.3. Video Maximal Marginal Relevance

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, it is different from the frames already selected in the summary. By analogy with the MMR algorithm, we define the Video Marginal Relevance (Video-MR) by:

$$\begin{aligned} \textit{Video-MR}(f_i) &= \lambda\ Sim_1(f_i, V \backslash S) \\ &- (1-\lambda) \max_{g \in S} Sim_2(f_i, g) \end{aligned} \qquad (5)$$

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_i is a candidate frame for selection. Based on this measure, a summary S_{k+1} can be constructed by iteratively selecting the keyframe with Video Maximal Marginal Relevance (Video-MMR):

$$S_{k+1} = S_k \cup \underset{f_i \in V \backslash S_k}{argmax} \begin{pmatrix} \lambda\ Sim_1(f_i, V \backslash S_k) \\ -(1-\lambda) \max_{g \in S_k} Sim_2(f_i, g) \end{pmatrix} \qquad (6)$$

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

- The average similarity is the arithmetic sum:
  $$AM(f_i, V \backslash S) = \frac{1}{|V \backslash (S \cup f_i)|} \sum_{f_j \in V \backslash (S \cup f_i)} sim(f_i, f_j) \qquad (7)$$
- The average similarity is the geometric sum:
  $$GM(f_i, V \backslash S) = \left[\prod_{f_j \in V \backslash (S \cup f_i)} sim(f_i, f_j)\right]^{\frac{1}{|V \backslash (S \cup f_i)|}} \qquad (8)$$

This leads to two variants: AM-Video-MMR and GM-Video-MMR. Both variants intend to model the amount of information that a new frame brings from the set of non-selected frames. GM is easily deteriorated by one bad factor, while AM is not. AM seems to be more stable. The better variant in these two would be selected only by their results in the experiment section 3.1.

Based on Video-MMR definition, the procedure of Video-MMR summarization is described as the following steps:

(a) The initial video summary S_1 is initialized with one frame f_1, defined as:

$$f_1 = arg \max_{f_i, f_i \neq f_j} \left(\prod_{j=1}^{n} Sim(f_i, f_j)\right)^{\frac{1}{n}} \qquad (9)$$

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.

(b) Select the frame f_k by Video-MMR:

$$f_k = arg\ max_{f_i \in V \backslash S_{k-1}} \begin{pmatrix} \lambda\ Sim_1(f_i, V \backslash S_{k-1}) \\ -(1-\lambda) \max_{g \in S_{k-1}} Sim_2(f_i, g) \end{pmatrix} \qquad (10)$$

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

(d) Iterate to step 2 until *S* has reached the desired size.

This algorithm has two variants, depending on which variant of the Video-MMR formula is used. Another issue is the value of the parameter λ, which can be used to adjust the relative importance of relevance and novelty. The next question is to select one variant and one λ, which are the best. This is explained in Section 3.1.

2.4. Comparison with user-made summary

Human evaluation is commonly considered as ground truth. So it is meaningful to compare Video-MMR to human choice. In a video set, 6 videos with most obvious features were chosen. Inside 6 videos, 3 videos own the largest distances with the others in this video set, while the other 3 videos have the smallest distances. The former represents the most unique contents, while the latter displays the most common contents of the video set. 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. This user-made summary is used as ground truth in experiment section 3.3. The reason of using 10 frames is that 10 keyframes are usually enough to present the contents of 1-5 minutes' video, and are easy for human to view and make the selection. 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 selections of a keyframe is 3, then $w = 3$. A keyframe that has never been selected by any user has a weight of 0. Similar to Eq. 3, the summary quality of Video-MMR with respect to the human choice can be defined as:

$$QC_{Video-MMR} = \frac{1}{m} \sum_{i=1}^{m} w_i \cdot max_{f \in S}\, sim(f, g_i) \qquad (11)$$

where *m* is the number of keyframes of the video set, and *f* is a frame of Video-MMR summary, *S*.

For further comparison, we also introduce the mean quality of every user-made summary compared with the other 11 user-made summaries:

$$\begin{aligned} QC_{human} &= \frac{1}{N} \sum_{n=1}^{N} QC_n, \\ where\ QC_n &= \frac{1}{m'} \sum_{i=1}^{m'} w_i \cdot max_{f \in S_n}\, sim(f, g_i) \end{aligned} \qquad (12)$$

In Eq. 12, $N = 12$, and m' is the unique keyframes' size of the other 11 user summaries, and frame f belongs to summary S_n. In this way, we can compare summary quality between Video-MMR, K-means and human choice (at least for a summary size of 10 keyframes).

3. EXPERIMENTAL RESULTS

This research is part of a joint project with an internet news aggregator web site (http://www.wikio.fr/). This web site gathers news items, (both texts and videos) from a large variety of sources, and organizes them by articles on specific topics. People can therefore have a global view of an event as presented through different channels and with various comments. For our experiments, we collected from their web site 88 sets of videos and an extra set "YSL", each set containing videos collected from various sources, but dealing with the same event. Every set includes 3 to 15 individual 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. Some video has the time length of more than 10 minutes, while some is only several seconds. And the genres of videos are very different, from advertisement, news, movie, and music to sports. These videos are used as one frame per second in the experiments.

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

$$sim\left(f_i, f_j\right) = \cos\left(H_{f_i}, H_{f_j}\right) = \frac{H_{f_i} \cdot H_{f_j}}{\|H_{f_i}\| \|H_{f_j}\|} \qquad (13)$$

where H_{f_i} and H_{f_j} are histogram vectors of frame $\mathrm{f_i}$ and $\mathrm{f_j}$. To define visual words, we first detect Local interest points (LIPs) in the image, based on the Difference of Gaussian and Laplacian of Gaussian, then compute a SIFT descriptor. The SIFT descriptors are clustered into 500 groups by K-means to compose a visual vocabulary with 500 words. The processing software to get visual word histogram is from [11].

This section is divided into four subsections: subsection 3.1 displays the summary reference comparison, which selects the best parameter and variant; subsection 3.2 compares Global Summarization (GS), and Individual Summarization (IS); subsection 3.3 shows summary qualities of Video-MMR by human evaluation. Furthermore, the comparison of Video-MMR, K-means and human summary is shown in subsection 3.3 too.

3.1. Summary Reference Comparison

For a given set of videos, we have to compare two possible variants of the algorithm and find the best possible value of the parameter λ. To select the best combination, we use the criterion of the minimum $d(S, V)$ according to Eq. 3. The minimum distance displays the most similar summary with original video set V, and then the parameters belonging to this distance are the required parameters of best Video-MMR. By sampling the possible values of λ into $0.1, 0.2, 0.3, \ldots, 0.9, 1.0$, we obtain a total of $2 \times 10 = 20$ combinations. This method is named as Summary Reference Comparison (SRC).

Fig. 1 shows SRC of AM-Video-MMR, whose summary size varying from 1 to 50 frames, and 10 curves are corresponding to parameter λ ranging from 0.1 to 1.0. Summary distances in Fig. 1 are the mean distances of 88 videos sets. Fig. 2 shows the same summary distances of GM-Video-MMR.

From Fig.1 and Fig. 2, we could conclude that the mean distance is globally the minimum, when $\lambda = 0.7$ and the variant is AM-Video-MMR. So this combination is the best for Video-MMR. And in the following experiments, we would use these values of the parameters.

3.2. Global and Individual summarization

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

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

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

We evaluate those summaries by computing their distances as Eq. 3 to the set of videos. We repeat experiments for different summary sizes. Fig. 3 shows an example of distances evolution for video set "YSL" whose summary sizes range from 1% to 15% of the size of the original video set. GS distance is substantially lower than the IS distance, so Global Summarization is preferable to the Individual Summarization. Following experiments use GS.

We repeated the experiment for all 88 video sets with summary percentage, 1%, 2%, 5% and 10% of original video. Most IS has larger summary distance than GS, and more than 85% of distance differences between IS and GS are less than 0.15.

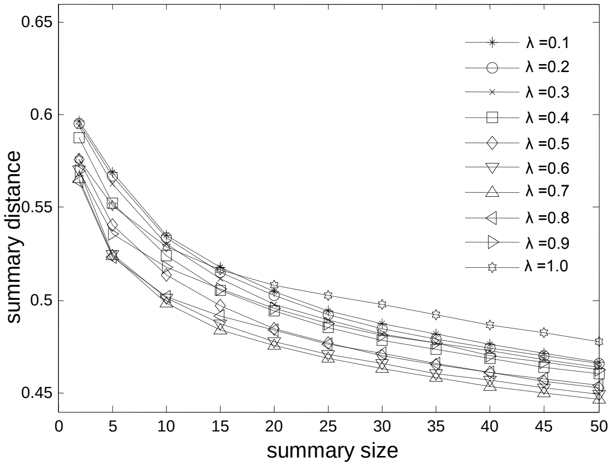

Figure 1. SRC of AM-Video-MMR

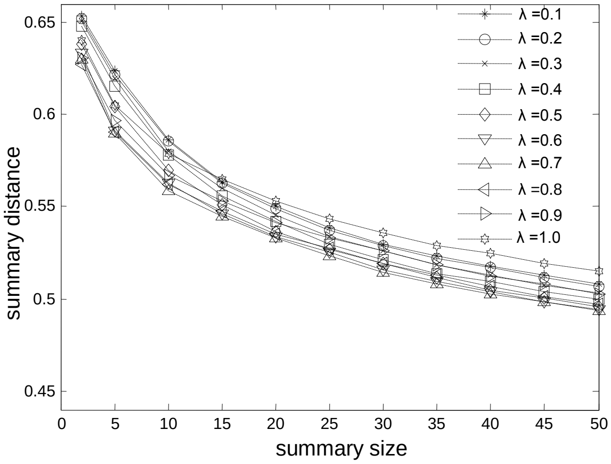

Figure 2. SRC of GM-Video-MMR

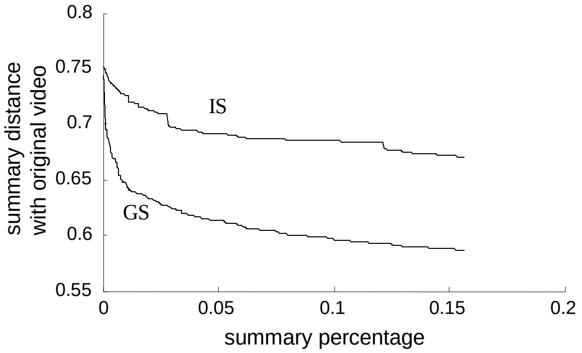

Figure 3. Comparison of "GS" and "IS" of video set "YSL"

3.3. Comparison of Video-MMR, K-means, and user-made summary

"YSL" video set, not one of 88 video sets, contains 14 videos. $QC_{Video-MMR}$ of Video-MMR summaries of "YSL" with 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 frames are shown in Fig. 4. We could see that $QC_{Video-MMR}$ increases with the summary size, because more similar information with human selection is included in Video-MMR summary. And $QC_{Video-MMR}$ monotonically increases with summary size, which proves that the summary quality of Video-MMR is stable.

Here we compare Video-MMR with K-means algorithm, which is a usual and popular clustering and summarization method. Summary quality of K-means is defined as $QC_{K-means}$ like Eq. 11 too. $QC_{K-means}$ evaluated by user-made summary is also shown in Fig. 4. It illustrates that summary qualities of Video-MMR are better than those of K-means, because the values of $QC_{K-means}$ are smaller than $QC_{Video-MMR}$. This is mainly caused by the random property of initial centers of K-means algorithm.

Compared with K-means, Video-MMR has some advantages:

(a) It avoids random initial centers, so it is more stable than K-means algorithm,

(b) It considers inter- and intra- relations of summary,

(c) It achieves dynamical summarization, which means that it could compute larger summary based on existent summary with smaller size. Video-MMR is more natural than K-means to do so.

Furthermore, QC_{human}, whose user-made summary from 6 videos of "YSL" is obtained by the method in section 2.4, is shown in Fig. 4. $QC_{Video-MMR}$ with size 10 is closer to this ground truth than $QC_{K-means}$.

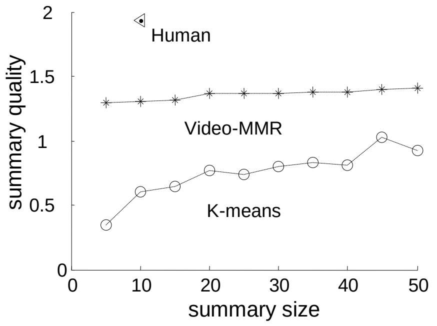

Figure 4. $QC_{Video-MMR}$, $QC_{K-means}$ and QC_{human}

4. CONCLUSIONS

In this paper, we have extended MMR text summarization algorithm to multi-video summarization. Two variants, AM-Video-MMR and GM-Video-MMR, have been identified. AM-Video-MMR experimentally shows to be better in two variants. Experiments also show that Global Summarization is preferable to Individual Summarization. At last, the summary quality of Video-MMR is evaluated by human results as ground truth. We also demonstrate that not only is the summary quality of Video-MMR better than K-means, but also it owns several advantages than K-means. In the future, we intend to improve Video-MMR and obtain better summary quality, by bringing more information into the algorithm, and explore further the specificities of multi-video summarization, in particular by considering global properties of video set, and on the aspect of evaluation.

5. REFERENCES

[1] Itheri Yahiaoui, Bernard Merialdo and Benoit Huet. 2001. "Automatic video summarization", *Multimedia Content-based Indexing and Retrieval,* Nantes France, 2001.

[2] Arthur G.Money. "Video Summarisation: A Conceptual Framework and Survey of the State of The Art", *Journal of Visual Communication and Image Representation*, Volume 19, 121-143, February 2008.

[3] Howard D.Wactlar, "Multi-Document Summarization and Visualization in the Informedia Digital Video Library", *Proc. of the 12th New Information Technology Conference,* Beijing China, May 2001.

[4] Emilie Dumont and Bernard Merialdo, "Automatic Evaluation Method for Rushes Summary Content", *International Workshop on Content-Based Multimedia Indexing*, London UK, June 2008.

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

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

[7] Kathleen McKeown, Rebecca J. Passonneau and David K. Elson, "Do summaries help? A task-based evaluation of multi-document summarization", *ACM SIGIR conference,* Salvador Brazil, August 2005.

[8] Lin, Chin-Yew, "ROUGE: A Package for Automatic Evaluation of Summaries", *In Proceedings of the Workshop on Text Summarization Branches out*, Spain, July 2004.

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

[10] Ryan McDonald, "A Study of Global Inference Algorithms in Multi-document Summarization", *Advances in Information Retrieval*, Volume 4425, 557-564, June 2007.

[11] http://vireo.cs.cityu.edu.hk (Video Retrieval Group, City U. of Hong Kong).