Static and Dynamic Video Summaries

Yingbo Li, Bernard Merialdo
EURECOM
Sophia-Antipolis, 06560, France
{yingbo.li, bernard.merialdo}@eurecom.fr

Mickael Rouvier, Georges Linares
University of Avignon
Avignon, 84911, France
{mickael.rouvier, georges.linares}@univ-avignon.fr

ABSTRACT
Currently there are a lot of algorithms for video summarization; however most of them only represent visual information. In this paper, we propose two approaches for the construction of the summary using both video and text. One approach focuses on static summaries, where the summary is a set of selected keyframes and keywords, to be displayed in a fixed area. The second approach addresses dynamic summaries where video segments are selected based on both their visual and textual content to compose a new video sequence of predefined duration. Our approaches rely on an existing summarization algorithm, Video Maximal Marginal Relevance (Video-MMR), and its extension Text Video Maximal Marginal Relevance (TV-MMR) proposed by us. We describe the details of those approaches and present experimental results.

Categories and Subject Descriptors
I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *video analysis.*

General Terms
Algorithms, Design.

1. INTRODUCTION
Video summarization has attracted a lot of attention from researchers these years, because of the unimaginable explosion of multimedia information. For example, the benchmark activity, the TREC Video Retrieval Evaluation (TRECVid), is important in the area of multimedia now. Many algorithms have been proposed to summarize single and multiple videos [2]. Some algorithms only depend on visual information [2], while others use visual and audio information [3], visual and text information, or all three kinds of information [4] [6]. The information used in the summarization algorithms may be diverse, but the summary itself is often built simply from the video frames [7].

A video summary can take two forms [5]: a static storyboard summary, which is a set of selected keyframes, or a dynamic video skim, composed by concatenating short video segments. According to their intrinsic properties, static summaries can contain video frames, possibly some keywords, but cannot include the audio track; while in dynamic summaries, all three kinds of information can be present.

In this paper we consider the construction of the static summary composed of keyframes and keywords. We assume that the display space has a fixed size, which has to be optimized between keyframes and keywords. A keyframe occupies more space than a keyword, but also generally contains more information. We search for an algorithm to optimally decide the percentage of keyframes and keywords that provide the maximum information inside the available display space. This allows building a static summary which contains the maximum information presented to the user.

For dynamic summary, we consider the synchronized summary, where the audio-visual segments are extracted from the original sequence and concatenated. We explore the issue of the optimal segment duration, since a short duration is generally sufficient to represent the visual content of a video segment, while a longer audio segment provides more information.

2. Linguistic information measure
In our approach, the information content of the audio track is evaluated based on the text transcription of the audio channel by an Automatic Speech Recognition (ASR) system from LIA (Laboratoire d'Informatique d'Avignon, France). The LIA ASR system is using context-dependent Hidden Markov Models for acoustic modeling and *n*-gram Language Models (LM). Training corpora comes from broadcast news records and large textual materials: acoustic models are estimated on 180 hours of French broadcast news. Language Models are trained on a collection of about 109M words, from French newspapers and large newswire collections. The ASR system is run on the audio track of the video sequences. The result is a sequence of words, with the beginning and ending times of their utterance. These timecodes allow synchronizing the audio and the video information in the summarization algorithm. They also allow providing candidate boundaries for audio-visual segments to be selected.

By analogy with text information retrieval techniques, the audio information content is measured according to the words that appear in the selected segment. We construct a word document vector d for the whole transcription of a video (or the transcriptions of a set of videos), as in the Vector Space model. We construct a similar vector for the text transcription t of a segment extracted from an audio-visual sequence. The audio information content of the segment is defined as the cosine between these two vectors:

$$sim(t,d) = cos(t,d) = \frac{t \cdot d}{\|t\| \|d\|} \quad (1)$$

The results are provided as lists of sliding windows of n words, (with n ranging from 1 to 10), together with windows covering complete sentences. For each window, the beginning and end times are provided, together with the similarity score. An example of such list for 3-grams is shown in Table 1.

Table 1. Some examples of 3-gram

Score	Begin	End	3-gram
0.06	51.53	52.18	on craint on
0.07	51.58	52.28	craint on s'
0.07	51.94	52.86	on s' exprimer
0.15	52.25	53.44	s' exprimer comédien
0.15	52.46	53.54	exprimer comédien ce
0.15	52.86	53.94	comédien ce matin
0.07	53.47	54.21	ce matin les

3. TV-MMR

3.1 Video-MMR

By analogy with text summarization, we have proposed to adapt the Maximal Marginal Relevance (MMR) [1] principle to design a new algorithm, Video-MMR [2], for multi-video summarization. When iteratively selecting keyframes to construct a summary, Video-MMR selects a keyframe whose visual content is similar to the content of the videos, but at the same time different from the frames already selected in the summary. By analogy with the MMR algorithm, we define the Video Marginal Relevance (Video-MR) of a keyframe at any given point of the construction of a summary *S* by:

$$\textit{Video-MR}(f) = \lambda\, Sim_1(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_2(f,g) \quad (2)$$

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. λ allows adjusting the relative importance of relevance and novelty. Sim_2 is just the similarity $sim(f,g)$ between frames f and g. And

$$Sim_1(f, V\backslash S) = \frac{1}{|V\backslash(S\cup f)|} \sum_{g \in V\backslash(S\cup f)} sim(f,g) \quad (3)$$

A summary S_{k+1} can be constructed by iteratively adding the frames with Video-MMR into the summary:

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

3.2 TV-MMR

The Video-MMR algorithm only uses visual information. In order to exploit the textual information obtained by the Speech Recognition, we propose an extension which we call Text Video Maximal Marginal Relevance (TV-MMR). TV-MMR selects video segments corresponding to *n*-grams by using both the textual and the visual content. By mimicking the formula of Video-MMR, the formula of TV-MMR is proposed as:

$$S_{k+1} = S_k \cup \underset{f \in V\backslash S_k}{argmax}\{ \beta[\lambda\, Sim_{I1}(f, V\backslash S_k) - (1-\lambda) \max_{g \in S_k} Sim_{I2}(f,g)] + (1-\beta)[\mu\, Sim_{T1}(f, V\backslash S_k) - (1-\mu) \max_{g \in S_k} Sim_{T2}(f,g)]\} \quad (5)$$

where f and g are audio-visual segments corresponding to *n*-grams. The definitions of Sim_{I1} and Sim_{I2} are the same as in Eq. 2. Sim_{T1} and Sim_{T2} are the textual similarities from ASR results, and they play a similar role for the text as Sim_{I1} and Sim_{I2} for the video. The parameter β allows adjusting the relative importance between visual information and textual information.

While in Video-MMR, the basic information unit was a single keyframe, in TV-MMR it is an *n*-gram segment. The visual content of an *n*-gram segment is composed of all the keyframes which appear between the beginning and ending times of the utterance. For faster computation, we subsample the video at the rate of 1 frame per second, so that a 5 second utterance will be represented by a set of 5 keyframes. The similarity between keyframes that is used in Video-MMR is extended to a similarity between sets of keyframes by computing the average of keyframes similarities.

The procedure of TV-MMR summarization is explained as the following sequence of steps:

1) The initial video summary S_1 is initialized with one segment, defined as:

$$S_1 = \underset{f_i, f_i \neq f_j}{arg\,max} [\prod_{j=1}^{n} Sim_I(f_i, f_j) \cdot \prod_{j=1}^{n} Sim_T(f_i, f_j)]^{\frac{1}{n}} \quad (6)$$

where f_i and f_j are *n*-gram segments from the video set V and n is the total number of segments except f_i. Sim_I computes the similarity of visual information between f_i and f_j; while Sim_T is the similarity of text information between f_i and f_j.

2) Select the segment f_k by TV-MMR formula, Eq. 5.
3) Set $S_k = S_{k-1} \cup \{f_k\}$.
4) Iterate step 2) until *S* has reached the predefined size.

4. STATIC AND DYNAMIC SUMMARIES

4.1 Static Summaries

A static video summary is basically composed of selected keyframes. However, it can be useful to use also some of the display space to show some keywords which are related to the content of the video sequence. In our work, we use the speech transcription of the audio track, as described in Section 2. The summary is often presented inside a display space with predefined size, for example a web page. Therefore, the summarization algorithm has to select a predefined number of keyframes to fit inside this space, while maximizing the amount of information which is presented to the user. When keywords are also possible, the summarization algorithm should decide, not only on which keywords to display, but also about the relative number of keywords and keyframes to fit in the predefined space. The diversity of the visual and the textual content is different from video to video, so that a fixed choice for the number of keywords and keyframes cannot be optimal.

In our work, we have considered that keyframes are of fixed size (another option would be to allow some keyframes to shrink, but we leave it for future exploration), and that the space occupied by a keyframe is equal to the space occupied by 60 characters. Selecting more keyframes reduced the number of words which can be displayed, and vice-versa. For a fixed display space, only combinations of keyframes and keywords which fit inside this space are considered. The task of the summarization algorithm is to find the combination that provides the most information.

Our video summarization algorithm, Video-MMR, is incremental, and produces a sequence of video summaries where one keyframe is added at each step. This provides a sequence of keyframes with decreasing visual importance, out of which we can easily consider the first *k*, for any value of *k*. During the Video-MMR, the marginal relevance $k_V(i)$ of a keyframe f_i as defined in Eq. 2, decreases as the iterations proceed. We fix a threshold and stop the Video-MMR iterations when the marginal relevance falls below the threshold. For a given video, this provides a number *M* of keyframes. We normalize the visual relevance of the keyframe:

$$k'_V(i) = k_V(i) / \sum_{j \in M} k_V(j) \quad (7)$$

From the speech transcription, we can associate each video keyframe with an *n*-gram, based on the timecodes. This allows defining the text similarity $k_T(i)$ of the text segment associated to the keyframe f_i as the cosine measure introduced in Section 2. Again, we normalize these values over the selected set:

$$k'_T(i) = k_T(i) / \sum_{j \in M} k_T(j) \quad (8)$$

We take the size of a keyframe as the basic unit, and assume that the available display size is *P* times the size of a single keyframe. As mentioned previously, size of a character is taken as 1/60 of the keyframe size. With these figures, the optimal summary will be composed of the set of keyframes ρ_V and the set of keywords ρ_T which maximize:

- The optimal summary to be presented in a display space best combination of frames and text is the one that maximizes the

total visual and textual information that is presented, as is described in the following formula:

$$max_{\rho_V,\rho_T}[K_V(\rho_V) + K_T(\rho_T)] \quad (9)$$

With the constraint $size(\rho_V) + size(\rho_T) \leq P$, and the definitions:

- $K_V = \sum_{j\in\rho_V} k'_V(j)$,
- $K_T = \sum_{j\in\rho_T} k'_T(j)$,
- $size(\rho_V) = |\rho_V|$,
- $size(\rho_T) =$ (number of characters of words in ρ_T)$/60$.

4.2 Dynamic Summaries

Our dynamic summaries are the concatenation of audio-visual segments extracted from the original videos. The candidate segments out of which we select are the segments corresponding to the utterances of *n*-grams. In this paper, we only discuss the dynamic summaries from the viewpoint of maximizing the information in summaries, though the story flow and rhythm are also important for the dynamic summary.

A specific difficulty comes from the fact that the rate of information flow is different between the audio and the visual media. For the visual part, videos are a succession of shots. Those shots are often rather long (on the order of 10 seconds or more), with slow motion (with the exception of music clips). In this case, a visual presentation of 1 or 2 seconds of the shot is sufficient to convey most of the visual content of the shot. Any longer presentation is a wasteful usage of the visual information channel for the summary. On the contrary, for the textual part, redundancy is extremely rare, so that longer extracts provide greater information content. Therefore, the choice of the optimal duration of *n*-grams is lead by two opposite constraints:

- Smaller values of *n* favor more visual content to be presented (for a given summary duration),
- Higher values of *n* allow more coherent text information to be included.

Based on this analysis, we explore the use of TV-MMR to find the best compromise between those constraints. For each value of *n*, we can build a summary from the *n*-gram segments. We can then compare the quality of these different summaries and select the best one according to a combination of its visual and textual content. We propose the following equation for this optimization:

$$argmax_n\, K(S_n) = argmax_n\left[K'_V(S_n) + K'_T(S_n)\right] \quad (10)$$

where S_n is the summary built by TV-MMR from the *n*-gram segments, $K(S_n)$ is the quality of its audio-visual content, defined as the sum of K'_V, the similarity of video segments in the summary with the original video and K'_T, the similarity between text words in the summary and all the text. Before applying TV-MMR, we define the expected duration of the summary.

We then perform experiments to compare the values of text similarities and visual similarities from different values of *n*, in order to find the best compromise.

5. EXPERIMENTAL RESULTS

In the experiments the video sets are collected from Internet news aggregator website "wikio.fr". Totally we have 21 video sets, each of which contains between 3 and 15 videos, whose durations vary from a few seconds to more than 10 minutes. The genres of the videos are various including news, advertisement and movie, to ensure the diversity of the experimental videos.

In the experiment, the similarity of two video frames, $sim(f_i, f_j)$, is defined as cosine similarity of visual word histograms:

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

where H_{f_i} and H_{f_j} are histogram vectors of frame f_i and f_j. And for the similarity of text of two segments in TV-MMR, it uses the same definition with Eq. 11 but the text histogram of an utterance is defined as:

$$H = (w_1, w_2, \ldots, w_T) \quad (12)$$

where w_T is the number of $T st$ word in the utterance, and the number of the words is T.

5.1 TV-MMR

To remain consistent with Video-MMR, we still use Summary Reference Comparison (SRC) in [2] to select the best parameters μ and β. First we vary μ from 0.1 to 0.9, each step being 0.1. Then we get a figure for 2-gram as the basic unit in Figure 1:

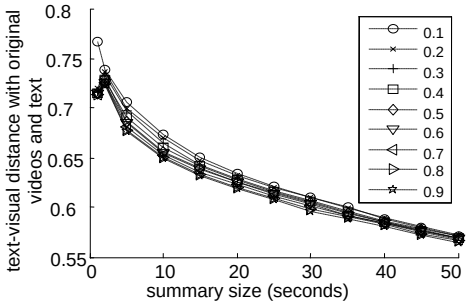

Figure 1. SRC of parameter μ

It is obvious that $\mu = 0.9$ is the best in Figure 1. For the other *n*-grams, the figures are similar with $\mu = 0.9$ owning the best curves, but they are not shown because of the limited pages. Therefore in Eq. 5 we prefer $\mu = 0.9$. And we vary β like μ and consider βs for different *n*-grams, finally we choose $\beta = 0.1$.

Because we have known $\lambda = 0.7$ in Video-MMR [2], in Eq. 5 $\lambda = 0.7$, $\mu = 0.9$ and $\beta = 0.1$. After the best parameters are decided, we can compare the text-visual distances with original videos of TV-MMR and Video-MMR in Figure 2. In Figure 2, we only show the examples of 2-gram and 8-gram, but the other *n*-grams have similar curves. It is obvious that our TV-MMR outperforms the existing algorithm Video-MMR.

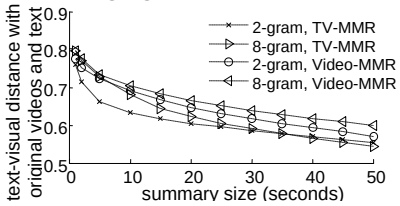

Figure 2. TV-MMR and Video-MMR

5.2 Static Summaries

For our experiments, we consider several display size:

- *P*=12, as a reasonable value when the display space is a full screen on a computer,
- *P*=6, a common value when using the display of a smart phone,
- *P*=3 and *P*=4, as often found when a single line of keyframes is considered, inside a larger page.

We perform experiments over 21 different video sets, representing more than 200 videos. For each set, we consider different values of $|\rho_V|$, select the corresponding keyframes and keywords, and plot the value of the total visual and textual information in the summary, as defined in Eq. 9. Figure 3 is the curve for the case where the display size is *P*=12, the text segments are 2-grams, and $|\rho_V|$ varies from 0 to 12. The maximum value is obtained for

$|\rho_V| = 5$. Table 2 shows the overall results of the optimal value of $|\rho_V|$ for various values of *P* and various lengths of *n*-grams. We can see that the optimal number of keyframes has little variations when different lengths of *n*-grams are considered. However, when full sentences are considered for the text segments, selecting a complete sentences force to select both important and unimportant keywords, which is suboptimal, and only keyframes are selected in the final summary.

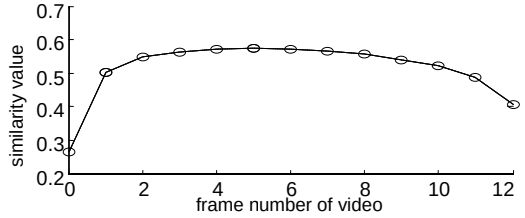

Figure 3. Information value of different $|\rho_V|$ when *P*=12 and $gram_n = 2$

Table 2. Statistical data of the best frame number in *P*

frame	1-gram	2	3	4	5	6	7	8	9	10	sentence
P=12	4	5	5	5	4	4	5	4	5	5	8
P=6	2	2	2	2	2	2	3	2	3	3	4
P=4	1	2	2	2	1	2	1	2	1	2	4
P=3	1	1	1	1	1	1	1	1	2	2	3

In Figure 4 we show an example of the static summary for $P = 6$ and 1-gram (For better visualization, the total space is not exactly 6 times the space of an image).

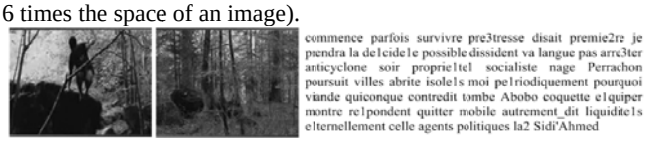

Figure 4. An example of the static summary

5.3 Dynamic Summaries

To obtain dynamic summaries with the duration *D*=10, 30, 50, 60, 70, or 80 seconds, we carry out TV-MMR with different grams, 1~10-gram or sentence, as the basic unit. Then we compute text similarities with utterance collection and visual similarities with the original videos for dynamic summary as Eq. 11 and Eq. 12. The mean text similarities and visual similarities of 21 video sets from different *n*-grams are shown in Figure 5. When the summary time is short like 10 seconds and 30 seconds, text scores don't increase with the increase of $gram_n$. However, when the summary time is around 60 seconds, $gram_n$ begins to influence on text similarity. The points of 7-gram are inflection and moderate points which maximize both text and video similarities for *D*=50, 60, 70, or 80 seconds.

Therefore, 7-gram is the best length of the basic unit/segment for dynamic summary, maximizing both text and visual information in a dynamic summary with a duration more than and around 60 seconds. A short basic unit, like 1-gram, seems to be better when the summary size is shorter than 50 seconds. According to our experimental data, the average durations are 2.1 seconds for 7-gram and 0.3 seconds for 1-gram. Therefore in dynamic summary of 60 seconds, every basic segment should last for 2.1 seconds.

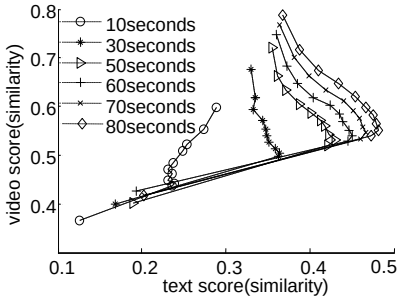

Figure 5. Dynamic summaries: the points from top to down are the values of 1~10-gram and sentences in each curve

6. CONCLUSION

We have proposed two strategies for maximizing the amount of audio, text and video information provided by a summary.

For static summaries we have presented a summarization algorithm which selects keyframes and keywords to maximize the visual and textual information presented in a predefined display space. Our algorithm automatically chooses the optimal number of keyframes. The visual and textual information of the candidates are evaluated, normalized, and the best selection is selected based on Video-MMR.

For dynamic summaries based on the concatenation of selected short video segments, we have proposed a novel summarization algorithm for text and video, TV-MMR, by which we decide the best segment duration by maximizing the summary information. With our models we can optimally construct dynamic summary of audio and video.

7. Acknowledgements

This research was partially funded by the national project RPM2 ANR-07-AM-008.

8. REFERENCES

[1] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *ACM SIGIR conference*, Australia, 1998.

[2] Yingbo Li and Bernard Merialdo. Multi-Video Summarization based on Video-MMR. *WIAMIS*, 2010.

[3] M. Furini and V. Ghini. An Audio-Video Summarization Scheme Based on Audio and Video Analysis. *IEEE Cosumer Communications and Networking Conference*, USA, 2006.

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

[5] B. Truong and S. Venkatesh. Video abstraction: A Systematic Review and Classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 3, No. 1, Article 3, February 2007.

[6] Changsheng Xu et al. Automatic Music Video Summarization Based on Audio-Visual-Text Analysis and Alignment. *ACM SIGIR*, Brazil, 2005.

[7] M. Furini, F. Geraci, M. Montangero. VISTO: VIsual STOryboard for Web Video Browsing. *CIVR*, 2007.