SEARCHING SEGMENTS OF INTEREST IN SINGLE STORY WEB-VIDEOS

Mickael Rouvier[1]*, Georges Linarès*[2]*, Benoit Favre*[4]*, Bernard Mérialdo*[3]

[1]LIUM - Univ. of Le Mans, [2] LIA-CERI - Univ. of Avignon, [3] Eurecom, [4] LIF - Aix-Marseille Univ.

ABSTRACT

This paper presents a method for predicting the parts of a video that could be marked as "interesting" by a user. Our approach consists in considering the three competing criteria : *salience*, *expressivity* and *significativity* and to automatically combine the three corresponding functions of interest. We evaluates this system on a user-annotated test set. Results demonstrate that, in spite of the intrinsic subjectivity of user choices, the system succeeds in finding about 51% of interesting segments.

Index Terms— audio processing, multimedia summarization, video categorization

1. INTRODUCTION

Low-cost video capture devices, Web 2.0 and video sharing platforms enabled users to easily acquire and share large amounts of multimedia content. Unfortunately, shared videos are of variable interest and crawling or searching relevant videos – or relevant segments – in these huge databases may be quite difficult.

Summarization may offer an efficient way of providing the user with compact views of videos. Most video summarization methods developed by the image processing community consist in picture extraction : abstracts are composed by a small set (typically from four to ten) of images extracted from the video. Numerous methods have been proposed for building such abstracts and this task was a track of a TREC evaluation campaign[1]. These methods generally do not take into account the audio channel, even if it could bring relevant information about content, in particular about the video semantics.

In this paper, we explore an alternative summarization paradigm that consists in extracting the most interesting segments from a video. These extracts are obtained by analysing jointly the audio channel and visual descriptors.

The extraction of segments of interest leads to two major difficulties related respectively to the segmentation (i.e. the search for segment boundaries) and to the characterization of the user's interest, that depends on the individual's sensibility, culture, emotional state, viewing context, etc. Some work, mainly from Humanities, studied the perception of videos by consumers and tried to characterize interest [**?**]. Our proposal is to consider three main criteria that may motivate user interest: salience, the expressivity and the significativity. We propose an estimation function of interest corresponding to each criterion, and a selection process that allows to predict the most relevant criterion given video content.

The other critical point for video summarization is segmentation. It is a classical issue of both video and speech processing, typically focusing on video frames, semantics, topics, etc. Here, the segmentation process aims at selecting a relatively small number of segments that may be interesting from the user point of view, considering both audio and video sources and respecting duration and continuity constraints. We propose an original algorithm to segment the video stream according to potential audio and video breakpoints.

The next section presents an overview of the segment-of-interest extraction process. Then, each of the three feature types are introduced and the criterion selection method is described. Section 3 introduces the segmentation process and present the algorithm. Section 4 presents the experimental setup and the results we obtained. Section 5 concludes and proposes some perspectives.

2. SYSTEM OVERVIEW

The system has to extract the best segment-of-interest (SOI) from a video. It relies on the assumption that one of the following criterion could motivate the user interest : *expressivity :*, which characterizes the form of the document rather than the content, the *salience :*, corresponding to unexpected moment, non-predictable, and *significativity :* that refers to the semantic content of the document.

Our proposal is to simulate such behavior by estimating a function-of-interest (FOI) for each of these criteria, and to train a statistical classifier to select the best FOI. When the FOI is chosen, a set of segments is selected by our algorithm.

3. CORPUS AND EVALUATION SETUP

The corpus is composed of videos selected from the web site Dailymotion[2]. We collected, during eight days, the 15 most-

Thanks to ANR agency, SUMACC Project for funding.

[1]TREC Video Retrieval Evaluation : http://trecvid.nist.gov/

[2]Dailymotion: http://www.dailymotion.fr

viewed videos, according to statistics provided by Dailymotion. Finally, we obtained a corpus composed of 120 videos with duration varying from three to five minutes.

The SOI of the collected videos were annotated by naive users, which had to select one SOI in each video, without any indication about how to motivated their choice. SOI duration had to be shorter than 30 seconds, but users could also indicate that there was no SOI in the video. 35 persons participated in the evaluation, each assessing an average of six videos.

Considering the fixed limit of SOI duration (<30s), the system is evaluated in terms of recall:

$$Recall = \frac{\text{Number of frames of SOI correctly detected}}{\text{Number of frames of SOI in reference}} \quad (1)$$

Performance is evaluated by measuring the intersection time of detected and reference SOI, with a frame every 10ms.

4. AUDIO AND VIDEO SEGMENTATION

Document segmentation is the first step for video summarization. Here, the objective is to identify the smallest meaningful segments that can be extracted and viewed out of their initial context.

This minimum size may be different depending on the task and document: for example, in audio documents the minimum size may be a speaker turn, sentence, or a topically coherent segment. Segmentation requires special care because an error in segmentation may impact largely the perceived quality of extractive summaries.

To reflect the diversity in potential segmentations, we propose to create a graph from audio and video segments. Then, the video summarization algorithm can search this graph for the optimal path according to specific duration and continuity constraints. We hypothesize that this best path corresponds to a selected SOI. The graph is built in 3 steps: (1) audio and video segmentation, (2) detection of connected nodes in the graph (connected nodes corresponds to adjacent video segments) and (3) merging of video segments that split an audio segment. This last constraint avoids the split of atomic audio segments, preserving minimal but meaningful parts.

The LIUM SpkDiarization speaker segmentation toolkit [1] is used for audio segmentation. Video segmentation is achieved by a scene segmenter provided by Eurecom.

5. MODALITY-DEPENDANT FEATURES

5.1. Extractive summarization algorithm

Summaries are generated by selecting a set of SOI from multiple videos under a duration or number of segment constraint. The video summarization algorithm aims at maximizing the global interest of the segment selection while minimizing overall redundancy of the information presented. The Maximal Marginal Relevance algorithm (MMR) is typically used in automatic text summarization. In automatic text summarization computational complexity problem is $O(e^n)$ where n is the number of sentence. But in our problem the SOI correspond to the agglomeration of several contiguous segment. So, the computational complexity problem is reduce to $O(n^2)$ where n is the number audio-video breakpoints. We propose to extract all contiguous segments (that does not exceed a durate of 30 seconds) and select the SOI that maximize the overall interest.

5.2. Significativity

In video summarization, a segment of interest (SOI) is a sequence which in the semantic content is significant. For example, a sequence evoking a news item which is the main subject of the video can be SOI.

The extraction of such sequence may be based on methods developed for text summarization. Here we follow the approach originally proposed in [2] which tries to extract a set of concepts and integrate them in the summary. To extract these concepts, we propose to use the method initially proposed in [3]. This method consists in extracting relevant word n-grams. The weight of a concept (modeled here by an n-gram) is based on the size of the n-gram and its IDF score:

$$w_i = n \cdot max_j idf(word_j) \quad (2)$$

where w_i is the weight of concept i, n is the length the $n-gram$ and idf the inverse document frequency score computed with frequencies from wikipedia. Moreover, the interest of semantic content may depend on the news. News can give us information on the relevance of the sentences to include in a summary. We propose to computing sentence importance according to their proximity to the news as it is posted on the Web. We use all the dispatches of news that have been selected by *Google News*. The weight of a sentence is computed as its cosine similarity with the closest web news item:

$$F^{significatity} = (1-\lambda)(\sum_x w_x) + \lambda(\sum_x cosine_{(x,web)}) \quad (3)$$

where w_x is the weight of all concept in the sentence x, $cosine(x, web)$ is the cosine score between segment x and the closest piece of news from the web. The λ parameter is used to balance sentence and concept based scores.

The algorithm select all continuous sequence of segments and select the one that maximizing the global interest.

Results obtained with this model are reported in the Table 1. Performance is evaluated by intersecting the selected segments produced by the system with the reference ones that were annotated by users (see Equation 1). By only relying on word n-gram concepts for detecting the segment of interest of

Table 1. The results obtained using the criterion of significativity.

	Concept	Web	Significativity
Results 0.44	0.37	0.42	0.38

a video, 42% of reference SOI time is retrieved. By combining concepts and the web-based weighing strategy, detection rate improve from 42% to 44%.

5.3. Expressivity

This criterion relies on the intuitive idea that discourse style and user interest may be correlated: interest could be suggested by expressive speech, in which the speaker involves his own sensitivity or emotions. It is important to note that here, the focus is not on the user (i.e. the perception of the video by users) but on the speech content in the video [4]. To compute expressivity, we combine two complementary indices related to lexical expressivity and speech spontaneity. The first one estimates the expressivity of the lingusitic contents of the speech, by using a lexicon that was annotated in terms of word expressivity. The "expressivity score" is computed by a cosine measure between the document word frequencies (as extracted by a speech recognizer) and the expressivity lexicon.

Spontaneity level is computed by combining acoustic as described in [5]. This system estimates a score corresponding to one of the following spontaneity levels: level 1 corresponds to prepared or read speech, level 2 corresponds to slightly spontaneous speech and level 3 corresponds to highly disfluent speech.

The spontaneity and emotion indices are combined into the following expressitivty-based SOI detector:

$$F^{spontaneity} = (1-\lambda)(\sum_x cosine_{(x,emotive)})+\lambda(\sum_x spont_x) \quad (4)$$

where $cosine_{x,emotive}$ and $spont_x$ represent respectively the lexical and speech expressivity indices of x. The λ parameter aims at balancing scores related to concepts and segments.

Table 2. The results obtained using the criterion of expressivity.

	Emotion	Spontaneity	Expressivity
Results	0.22	0.14	0.34

The results obtained by using expressivity features are reported in Table 2. Results show that this criterion seems significantly less effective than the previous one, the expressivity-only system reaching 34% of good detection. Even though this criterion is not as good at predicting SOIs, we hope that it is complementary and will help in the combination system.

5.4. Salience

A part of a video could be interesting because something special and unexpected occurs in it. This point of view is strongly opposed to significatitvity since it evaluates how a segment is a good representative of the whole content. By opposition, salience indicates how much the segment is different form the rest of the video.

Technically, salience detection is close to novelty or model breaking detection, two topics from the automatic classification field that were largely explored in various contexts.

Typically, detection of salient segments first relies on modeling the whole document. Then, model breaking detection is achieved by comparing this background model to a local model estimated on a short term window.

Here, features involved in video characterisation are both form the audio and video channels. Video frames are represented according the bag of visual word model (bovw) [6]. This modeling paradigm consists in representing the document by pattern frequencies with an *a priori* set of representative patterns.

In our system, we use SIFT descriptors clustered in 500 classes as the bovw dictionary. Then descriptors are extracted for each frame of the video and we can compute segment-level histograms of the detection of each descriptor. This leads to the following visual salience estimator:

$$video(D_1, D_2) = 1 - \frac{\sum_i w_{1i} \sum_i w_{2i}}{\sqrt{\sum_i w_{1i}^2}\sqrt{\sum_i w_{2i}^2}} \quad (5)$$

where D_1 is the bovw histogram of the segment and D_2 the histogram for the whole video.

The audio model is based on a GMM (Gaussian Mixture Model) estimated on MFCC (Mel Frequency Cepstrum Coefficients) feature vectors. This is a statistical model that allows to estimate the likelihood of a segment knowing the whole video. This likelihood is naturally an index of the predictability of the segment considering the background model. The two estimators are combined to estimate a saliency score :

$$F^{salience} = (1-\lambda)(\sum_x image_x) + \lambda(\sum_x son_x) \quad (6)$$

where $image_x$ and son_x correspond respectively to the video model breaking and audio model breaking for segment x. The λ parameter is used to scale the relative importance of visual and audio scores.

Table 3 presents results obtained by this saliency-based detector of SOI. By using audio or video channels only, we obtain respectively 34% and 28% of correct detection. The audio-video combination clearly improves theses detection rates to 38%.

Table 3. The results obtained using the criterion of salience.

	Video	Audio	Sallience
Results	0.34	0.28	0.38

6. NATURE OF INTEREST DETECTION

Each of the three detectors previously described offer a complementary point of view of moments of interest in a video. The classical approach to make the best benefit from multiple views consists in combining them – a large variety of combination schemes have been proposed in the past. Combination usually relies on the assumption that a document may be viewed as a combination of latent factors. The case of user interest seems radically different: features correspond to exclusive factors, even opposite ones: saliency and significatibvity thrive in opposing directions, even if each of them may have related to expressitvity.

Considering that such a combination does not match to the fundamental nature of our descriptors, we choose a selective approach, where the interest of a segment is related to only one of the three detectors. Therefore, we propose a selection mechanism that choses the most relevant function of interest for a given video, from significativity, expressivity and salience. The selection mechanism relies on a function-of-interest-independent representation of segments and an automatic classifier trained to selecting the best FOI. The feature set is composed of eight indices mainly based on document structure, including speaking time, number of speakers, number of video frames, popularity of spoken contents (estimated by frequency analysis on news web sites), audio energy (mean, lower and upper bounds). These features are fed to an SVM classifier reranker trained to predict which FOI should be applied for a given video. Training data is created by running each detector on a development set and labeling each instance with the FOI that performs best in term of reference SOI time recall. Considering the small amount of training data available, the SVM is trained with a leave-one-out strategy: the corpus is split in eight subsets, the classifier being trained on seven of them and evaluated on the remaining one.

Table 4. Performance by using a SVM classifier.

	Signific.	Express.	Sallience	Classif.
Results	0.44	0.34	0.38	0.51

Correct classification rates reach 51% of SOI (according to our metric, which focuses on the coverage of reference segments of interest) in Table 4.

In order to estimate the potential of this selective approach, we compute in Table 5 an oracle score that represents the score we could achieve knowing the best FOI. The oracle correct classification rate is 68%. Therefore, our results are significantly worse that this oracle (about 10% relative). Nevertheless, this oracle demonstrates the complementarity of the detectors: an optimal combination could improve the system from 51% to 68%.

Table 5. Compare SVM classifier and oracle.

	Classif.	Oracle
Results	0.51	0.68

7. CONCLUSIONS

We presented a multi-modal approach for segment-of-interest extraction from Dalymotion videos. The global approach we followed consists in simulating user behavior to select the best segment of interest in a video according to significativity, expressivity and salience criteria. A reranker is trained to select the best detector given global video features.

Our experiments are conducted on a set of 120 videos that were annotated by naive users. These preliminary results validate the principle of the proposed approach: even though performance of single-FOI systems vary from 28% to 44%, their combination though the reranker yields a strong improvement of the segment-of-interest detector to 51%.

8. REFERENCES

[1] Sylvain Meignier and Teva Merlin, "Lium spkdiarization: An open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[2] Daniel Gillick and Benoit Favre, "A Scalable Global Model for Summarization," in *Human Language Technology conference (HLT-NAACL)*, 2009.

[3] Shasha Xie, Benoit Favre, Dilek Hakkani-Tür, and Yang Liu, "Leveraging sentence weights in concept-based optimization framework for extractive meeting summarization," in *Conference of the International Speech Communication Association (InterSpeech)*, 2009.

[4] Jackson Liscombe, Giuseppe Riccardi, and Dilek Z. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Conference of the International Speech Communication Association (InterSpeech)*, 2005.

[5] Mickael Rouvier, Richard Dufour, Georges Linars, and Yannick Estve, "A language identification inspired method for spontaneous speech detection," in *Conference of the International Speech Communication Association (InterSpeech)*, 2010.

[6] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision (ICCV)*, 2003, pp. 1470–1477.