Video Summarization at Brno University of Technology

Vítězslav Beran, Michal Hradiš, Adam Herout, Stanislav Sumec, Igor Potúček, Pavel Zemčík, Josef Mlích, Aleš Láník, Petr Chmelař
Brno University of Technology
Faculty of Information Technology
Department of Computer Graphics and Multimedia
Božetěchova 2, 612 66 Brno, CZ
{beranv, herout, sumec, potucek, zemcik, chmelarp}@fit.vutbr.cz
{xhradi05, xmlich02, xlanik00}@stud.fit.vutbr.cz

ABSTRACT
This paper describes the video summarization system built for the TRECVID 2007 evaluation by the Brno team. Motivations for the system design and its overall structure are described followed by more detailed description of the critical parts of the system, which are feature extraction and clustering of frames (shots, sub-shots) in time domain. Many ideas were not included into the system because of the time constraints. Those considered promising are stated and briefly described as possible future work.

The results of video summarization presented in this paper can be considered to be a humble success and can encourage further development in the field. This is specifically true as not all the features that can be considered and processing methods were implemented in the evaluated system.

Categories and Subject Descriptors
I.5.3 [**Pattern recognition**]: Clustering

General Terms
Algorithms, Similarity measures.

Keywords
Video, summarization, image features, time compression, TRECVID evaluation.

1. INTRODUCTION
Contemporary technology makes possible to acquire huge sets of video content e.g. from TV broadcasting, meeting rooms, security systems etc. Such data can be further reused for various purposes. However, searching of desired information within large video libraries is time consuming. It becomes necessary to give users summarizing and skimming tools, which allow speeding up this process. These tools should produce shortened versions of source videos with regard to the information content.

Various methods for creating of summarizing videos have been already proposed. One class of techniques is based on time compression. The playback rate of audio and video is speed up with almost no pitch distortion. However, these techniques are limited to relatively low saving factor around 1.5 – 2.5 depending on speech speed. Slightly better results can be achieved when silent intervals are completely removed. Different techniques generate a static storyboard of images which are selected according to information contained in video or audio tracks.

This paper describes the system for creating video summaries based on an identification of similar clips. The best representative clip from every group is selected and inserted into the final video. Further, the resulting summary is formatted with additional information, which helps to localize other occurrences of presented clip.

2. SYSTEM OVERVIEW
Different purposes of the resulting videos would call for different summarization methods. The presented work targets summarization for professionals who need to deal with a number of relatively long video records. The resulting video should then cover parts of the original recording, representing preferably all different flavors of shots. Therefore the resulting video is not supposed to contain the most interesting scenes, the most dynamic ones, or those with closest relationship to the "story", etc. Also the selected approach does not take into account any understanding of the semantic meaning of the separate shots. Automatic semantic understanding is at the moment not possible and the system is supposed to work for unknown videos, where some semi-automatic or guided approach would be possible.

The scheme of the video summarizing system is on Figure 1. The input video frames are described using preferred image features and classified – shot boundary and wanted/unwanted frame. The rough video is divided into short shots that are described and classified similarly as frames and finally clustered. Representative shots are combined to the final video according to the layout setup.

With the targeted purpose in mind, notable effort was invested in the resulting video layout – the output is not simply a sequence of (shortened) shots of the original video, but the actual video is played in a (though large) window, and is accompanied by textual

Conference'04, Month 1–2, 2004, City, State, Country.

and graphical information. The layout is described further in detail.

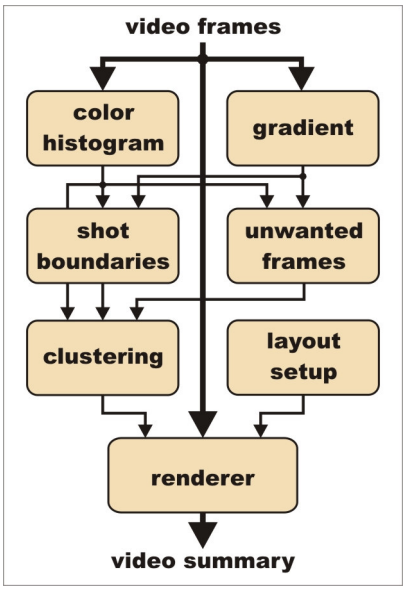

Figure 1 Schema of the system for video summarization.

The whole system was not developed with the intent of immediate use in practice by an end user. Therefore, the software architecture of the system is rather loose so that it allows experimenting and development. Much of the system (definitely the overall process control) is done by relatively simple batch scripts which are running relatively simple C/C++ binaries to perform specialized tasks on given data. This concept partly complicates collecting of the total system runtime – parts of the system operation were run separately one from another.

3. ALGORITHMS

The system is based on several basic image feature descriptors such as color histogram or image gradient distribution [1]. These features were input to a clustering algorithm, which selected "representative shots", which were included in the output video, arranged into the output "layout".

Figure 2 Example image used for descriptor visualizations.

3.1 Features

In the beginning, we made up several image descriptors based on different image features: color histogram, gradient distribution, Hough transform, motion vectors or simple texture analysis. The distinctiveness of all descriptors was evaluated on the development data and only two descriptors were chosen for final system: color histogram and gradient distribution (see Figure 3).

The color histogram descriptor uses image in HSV color model and computes the histogram in HS space. To improve the descriptors robustness, we divided the image into several parts and described sub-images separately.

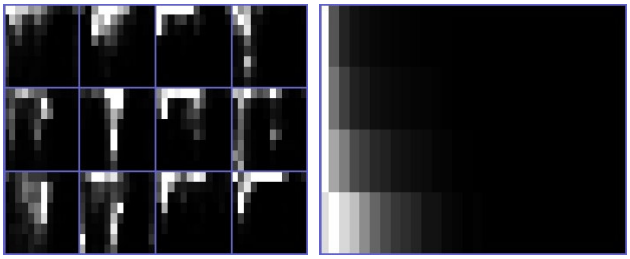

Figure 3 Visualization of Color Histogram (left) and Gradient descriptors (right).

The gradient distribution descriptor is computed from a histogram of the magnitude of the image intensity gradient. The gradient is computed on different scales so also low-frequency structures contribute to final description.

3.2 Shot Boundaries and "Unwanted Frames"

First of all, the term 'unwanted frames' should be better specified. All kind of images that should not appear in summarized video, e.g. color stripes or one color images, are unwanted frames represented by unwanted descriptors. It means that also frames around video shot boundary are unwanted. Detection of unwanted images is important for our clustering approach.

The set of unwanted descriptors has been made from several sets of frames with unwanted content. Frames with similar content type were described and averaged together composing the unwanted descriptor of one frame type. During video processing, decision, whether particular frame is 'unwanted', is based on minimal distance between the frame descriptor and all unwanted frames (Euclidean distance is used). When the minimal distance is bellow some threshold the frame is marked as unwanted.

Although we do not need the shot boundaries for video summarization, the acquaintance of them significantly improve the final result. The final summarized video is composed of short shot picked up from original video (for more detail see next chapter). Without knowing of shot boundaries, it could happen that final short shot would contain the cut. Our shot boundary detection is based on descriptor's changes. The derivation is computed using different window sizes, so both fast and slow cuts are covered.

3.3 Choosing Representative Shots

The original video is divided into possibly overlapping shots of constant length. For the submitted system, we used sequences with 2 seconds duration with 1.84s overlap. Each of the shots is assigned a feature vector which is formed by concatenating means and standard deviations of the feature vectors describing the frames in the shot. These shots are further treated as atomic units and directly represent the shots which form the final

summarization. To assure that none of the unwanted frames get into the summarization, we simply discard those shots which contain any unwanted frames.

We use PCA, which is computed separately for each of the original video sequences, to reduce data dimensionality. In the transformed feature space, we cluster the shots using k-means algorithm with Euclidean distance. The k-means algorithm is executed repeatedly with random starting conditions to increase probability of receiving optimal solution. Since the final summarized video is formed of single representative of each of the clusters, the number of clusters is chosen according to the length of the shots and a desired length of the summarized video. We choose such representative shots which are nearest to the center of its cluster. The representative shots are ordered in the summarization according to position in the original video.

We chose to split the video into shots of equal length for two reasons. First, we did not aspire to capture dynamic events. Instead, we wanted to present all distinct scenes from the original video for which the equal length of shots is suitable. This way an optimal length with respect to the speed of perception of the user can be used. Second, the length of the resulting video can be well controlled by choosing the number of clusters. Thanks to the high overlap of the scenes we do not loose almost any material around cuts in the video.

4. LAYOUT

As can be noticed from our summary video layout (see Figure 4 and Figure 5), we perceive the layout design as the crucial issue when composing efficiently quickly understandable video summary.

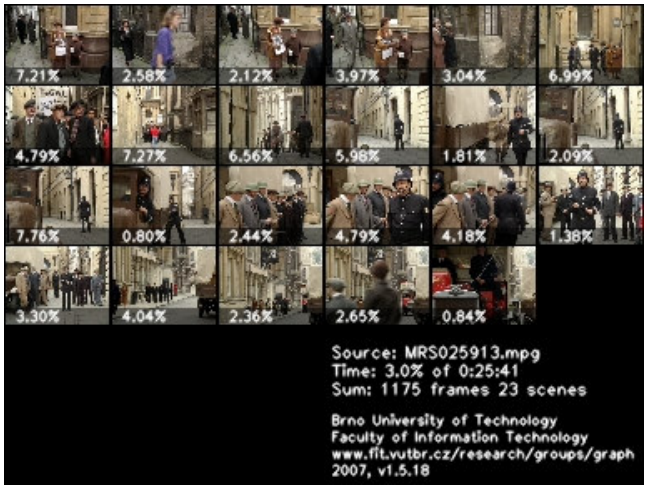

Figure 4 Frame with the overview of the summary video content.

One way how to introduce the video content to the viewer is to build up the contents. The thumbnails represent the significant scenes together with some more information about the scene. The information we used is the scene occurrence in the original video. We also included some statistical information such as the ratio between original and summary video, number of significant scenes or number of frames.

The summary video is built up from several short shots of the significant scenes. Because the original video may contain the significant scene several times, and on different parts, we used the timeline to depict the frames with the similar content to the particular scene. The main part of the layout is reserved to significant scene representative. Bellow the timeline, there is the strip with other representatives of the same significant scene but from the different parts of the original video. Each representative points to the timeline for better orientation where it comes from within the video. The rest of the layout is left for statistical information such as the number of the actual scene, the scene occurrence in the original video, the video name, etc.

Figure 5 Layout of the summary video.

5. RESULTS

The results achieved in TRECVID 2007 [4] evaluation are shown in Table 1 (note, please, that the table states also the rank of the methods after significance test - dependent t-test with significance level 0.05). Fairly good fraction of inclusions was achieved – the 12th rank is exactly in the middle of participant field.

Table 1 Ranks achieved at TRECVID 2007 evaluation and the range of ranks with significance testing.

Meassure	Rank	With significance test
Time	8	
Judge time	22	
Inclusion	12	9-14
Understandability	20	20-22
Redundancy	23	14-23

This suggests that the clustering approach is reasonable especially given the fact that not all the considered features were implemented. The fact that redundancy results were not too highly evaluated was mainly caused by the fact that the summary represents the video through a series of relatively short shots and that the clustering method did not adopt all the features that can be considered. The understandability of the summaries should be further improved by reduction the redundancy and improvements in the video and screen layouts.

6. FUTURE WORK

The relatively tight schedule of the evaluation did not allow many ideas to be incorporated into the final solution. Also the system was built (for the particular purpose) from the scratch, based only on low-level existing parts, such as some feature extraction routines etc. The work then concentrated into integration of the system and many trivial but necessary tasks. Having the base system set, the improvements would now become interesting and could lead to improvement of the system's performance. Let us mention some ideas that were considered for the system but were not included, generally for the time constraints.

One of the most crucial things is to analyze whether the source video is suitable for summarization and what is the best way to do it. There can be several different types of videos and each particular video needs quite different summarization approach. We hope that some analysis of shots content distribution might help to choose the best summarization algorithm and evaluate the length of the final video. Among several approaches that might help to improve the overall system performance, we are thinking of using features describing repetitive changes of patterns such as waves, smoke, fire, flag in the wind, a moving escalator, etc. The dynamic texture techniques seem to be promising.

Other possible algorithm extension is fomating of the output video according to specific aesthetical aspects. Some elementary rules, which describe shot composition, can be applied to achieve better visual impression. This option becomes quite important, if summary videos are targeted to the ordinary viewers. We plan to addapt existing rule based system that is designed for an automatic video editing of meeting data [2].

We also want to search for more suitable clustering methods. Clustering with Gaussian mixture model [3] seems promissing as it is able to fitt to the data well. Further, we could adjust the scene lengths as a postprocess step according to the dynamics of their content while keeping the current approach with equal sizes of shots.

7. CONCLUSION

This paper presented the solution for the rushes summarization task of TRECVID 2007, as it was developed by the Brno University of Technology team. Starting the summarization system from scratch, only based on existing classification and object-detection in-house software, the final results can be considered a humble success. An important thing is that the creation of the system induced many ideas for future development of the summarization engine and also brought some potential practical uses.

The system (mainly due to time constraints) remained very simple – consisting basically of per-frame feature extraction and following clustering of shots or groups of frames – which surprisingly did not handicap it too severely. The layout of the output (summarized) video is expected to be of use in practical applications – for a professional archiving or manipulating with video sequences. However, it did not appear to be helpful for the purpose of the evaluation/competition, or could have been even contra-productive.

A question is raised from the observations of the final results, whether any system not evaluating the semantics of the scenes could perform significantly better than the simple clustering of basic per-frame features. If not, the summarization engines should rely greatly on understanding the scene and the whole task is remarkably redefined.

The authors would like to express their warm thanks to the TRECVID 2007 organizers for taking the effort of preparing the evaluation and evaluating the results.

8. ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Education, Youth, and Sports centre of basic research "Centre of Computer Graphics", LC06008 and research project "Security-Oriented Research in Information Technology" CEZ MŠMT, MSM 0021630528, EU IST FP6 projects "AMIDA" EU-6FP-IST, IST-033812-AMIDA, and "CARETAKER" EU-6FP-IST, 027231. The authors would like to express thanks for all of the support.

9. REFERENCES

[1] Shirley, P. et. al. Fundamentals of Computer Graphics. *AK Peters, Ltd., 2nd edition*, ISBN 1-56881-269-8, 2005.

[2] Sumec, S. Multi Camera Automatic Video Editing. *In Proceedings of ICCVG 2004.* Warsaw, PL: Kluwer, 2004, 935-945.

[3] Dasgupta, S. Learning Mixtures of Gaussians. *Proc. of Symposium on Foundations of Computer Science (FOCS),* 1999.

[4] Over, P., Smeaton, A.F. and Kelly, P. The TRECVID 2007 BBC rushes summarization evaluation pilot. *In Proceedings of the TRECVID Workshop on Video Summarization (TVS'07).* Augsburg, Germany, September 28, 2007, ACM Press, New York, NY, 2007, 1-15.