Video Summarization at Brno University of Technology

Vítězslav Beran, Michal Hradiš, Pavel Zemčík, Adam Herout, Ivo Řezníček
Brno University of Technology
Faculty of Information Technology
Department of Computer Graphics and Multimedia
Božetěchova 2, 612 66 Brno, CZ
{beranv, ihradis, zemcik, herout, ireznice}@fit.vutbr.cz

ABSTRACT
This paper describes the video summarization system built for the TRECVID 2008 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. Low-level features, which are extracted from each frame, are clustered to group visually similar shots together. The final video summary production is an iterative procedure, where the probability, speed and trimming of each cluster candidate are evaluated until some criteria, such as final summary length, are fulfilled. The paper also contains the discussion about appropriate layout of the final video summary, taking into account experiences from the last TRECVID evaluation. The final conclusion points out the weak and strong aspects of the presented approach reflecting system performance in comparison with other state-of-the-art systems.

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

General Terms
Algorithms

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

1. INTRODUCTION
Contemporary technology makes it 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.

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 the 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 should not necessarily contain the most interesting scenes, the most dynamic ones, or those with closest relationship to the 'story', as often interpreted when we discuss 'video summarization'. Also the selected approach does not take into account any understanding of the semantic meaning of the separate shots. Fully-automatic semantic understanding is currently not achievable, so the system, that is supposed to work for unknown videos, would be possible only using semi-automatic or guided approach. The schema of the video summarizing system is in Figure 1.

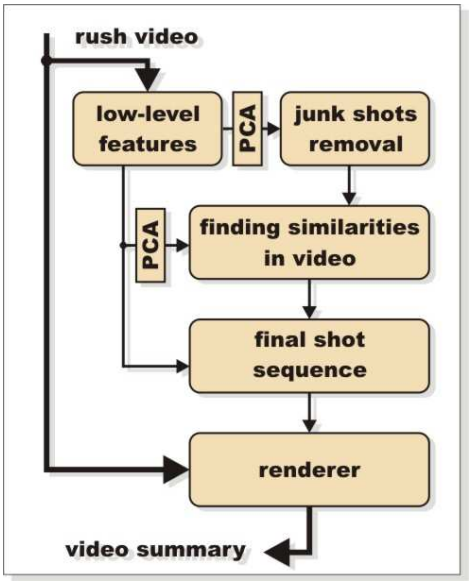

Figure 1 Schema of the system for video summarization.

The input video frames are described using low-level image features. Shot boundary detection did not prove to have much

influence on the final video summary, so we avoided this step in the system. Instead of describing the video shots itself, we work with video segments of predefined length (e.g. 1 second). The junk shot removal is based on manually annotated data from the training set that were segmented and results into several junk shot representatives. Each input frame is then also classified as a junk/non-junk frame. All valid shots, not containing junk frames, are clustered to produce information about similarities within the entire video. We have designed a metric to combine shot similarity, variability, length, etc. and evaluated the probability of the shot candidate, its speed and trimming parameters.

With the targeted purpose in mind and feedbacks from the approach carried out last year, notable effort was invested in the resulting video layout. In comparison with previous design [4], we have simplified the layout so that it contains only play speed information and a timeline.

3. ALGORITHMS

The system is based on several basic image feature descriptors such as color histogram and image gradient distribution [1]. These features served as input to a clustering algorithm, which selected "representative shots", which were then 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 the final system: color histogram and gradient distribution. The descriptors shown in Figure 3 are computed on the image on Figure 2.

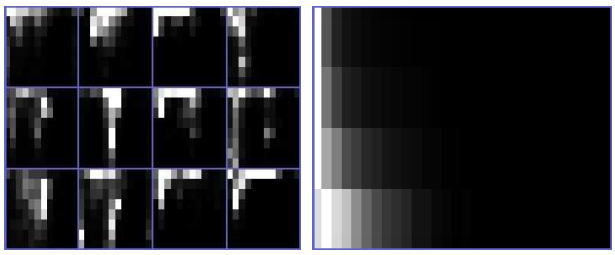

Figure 3 Visualization of Color Histogram (left) and Gradient descriptors (right).

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. The visualization of the color histogram descriptor is showed on Figure 3 (left) where each rectangle represents histogram of the sub-image values in sub-sampled HS space. The frequency of image subdivision and HS space quantization are input parameters for the method.

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 the final description. The Figure 3 (right) visualizes the values of histograms where rows represents different scales.

3.2 Junk Shots Removal

The junk shots are those video subsequences or frames which contain supportive, calibration and suchlike information (e.g. color stripes, one color images or clapboards, see Figure 4). The PCA method was used as the first preprocessing step to reduce data dimensionality. For the purpose of junk shot removal the PCA was computed over all development data. Then we manually annotated the video parts containing junk shots.

Figure 4 Examples of junk frames.

The model of junk shots is based on segmentation of junk frames features. K-means algorithm segments the junk frame features into several clusters that represents the junk frames to remove. Having the junk frame representatives we classify each input frame using Euclidean distance to the closest junk representative as an error measure. When the error exceeds predefined limit, the frame is classified as the junk frame. The limit error value for each particular representative was manually determined on all development data during segmentation process.

The junk shots removal also includes removal of the shots whose content would not cause their rejection but that are too short to contain any reasonable information.

The situation that the shot is rendered too short may occur during the initial stages of video evaluation, during definition of segments selection of representative candidates, or during final shots sequence building.

3.3 Finding Similarities in the Video

Clustering was used to find similarities in the video sequences. In the selected approach, PCA method is first computed separately for each of the original video sequences, to reduce dimensionality of the data. In the transformed space, the data is fit with a Gaussian mixture model (GMM) using the expectation maximization algorithm. The number of mixtures which are used is determined according to the desired summary length and the approximate average length of a single scene in the summary. The optimal number of the mixtures can be estimated only approximately, as the scenes in the final summary occupy variable time interval and some mixtures can be even completely

discarded. The expectation maximization algorithm is executed repeatedly with random starting conditions to increase probability of receiving the optimal solution. The optimality of the created models is estimated based on the model's likelihood over the data and a coherence measure C:

$$f(x) = \arg\max_{m} \left[P\left(GM_m \middle| x\right) \right]$$

$$C = \frac{1}{T} \sum_{t=1}^{T-1} \left\| f(x_t) = f(x_{t+1}) \right\|$$

where x_t is the feature vector representing t-th frame, T it the total number of frames and $P\left(GM_m \middle| x_t\right)$ is the probability that the feature vector was generated by Gaussian mixture m. More precisely, the models are assigned likelihood ranks and coherence ranks and the model with highest sum of these two ranks is considered the best and is used in the subsequent process. The result of this part of the summarization process is the likelihoods given by each of the best model's mixtures for each of the video frames.

The junk shot frames do not enter the similarity search at all. They are not used to compute the PCA, nor the expectation maximization. To propagate the information about the junk shot frames, these frames are assigned likelihoods of zero value.

After the original frames are marked using likelihood with which the frames belong to certain cluster, they are subdivided into shots that belong to certain clusters. The subdivision algorithm tolerates seldom occurrence of frames likely to belong into a different cluster; however, the pre-defined error measure should not be exceeded.

When the shots are defined, selection is performed. The criteria for the selection are:

- "Variability" inside the shots – a measure of the change of the features inside the shot. The shot with the highest variability is selected in the hope that the variability signals amount of information in the shot.
- Length of the shots – the longer shots may contain more information and also can be better time-compresses in order to produce appealing video output.
- At least one representative of each cluster should be preserved in the output if at all possible.

The above criteria are applied in an iterative process whose result is a sequence of shots that does not exceed a pre-defined time limit.

3.4 Final Shots Sequence

The final shots are prepared for the video summary production by adjustment of their speed and possibly through their trimming. The procedure of preparation relies on the following criteria:

- The "variability" per time in the video shots should be made constant. This is achieved through speeding up the shots with low amount of variability.
- The speedup is limited to pre-defined limit so that the video shots are not shown in unacceptably high speed that would prevent their proper understanding by humans.
- The size (time) of each of the output shots should not exceed a pre-defined limit.

After application of the above criteria, the total length (time) of the output video summary can be determined. If the total length exceeds the pre-specified limit, the parameters of the above criteria are modified and the process iteratively repeats till the total video length is below the pre-specified limit.

4. LAYOUT

We perceive the layout design as a crucial issue as it contributes on how efficiently and quickly the viewer can understand the video summary. Our previous effort ends up with video summarization layout presented in [4]. Our previous aim was to offer such information, that anyone viewing the summary is able to find the video source or precise position of desired scene. Also the information about actual position in the summary video, its lengths and scene occurrence was thought as necessary. Such solution turned out not to be completely understandable and efficient. It took some time to acquaint the viewer with the layout to use it efficiently.

When designing the new layout, we focused on entirely different goals than in the previous approach. The aim is to provide the user with only absolutely necessary information such as actual position in the video, positions and resemblance of the similar shots to the actual one and the speed of the actually played shot. The new layout is shown in Figure 5.

To represent the introduced information in an understandable way we used the timeline structure. The currently presented shot is emphasized by red color. The similarities of all other shots to the current one are represented by intensity on the timeline. The more similar the source shot is the brighter is the timeline. The only textual information displayed to the viewer is the relative play speed.

Figure 5 Video summary layout.

5. RESULTS

The results of the presented approach to video summarization were evaluated in the TRECVID 2008 [4] evaluations (see Table 1). The results show that the created summaries have relatively pleasant tempo and that they are relatively easy to understand (according to the judge time). The summaries contain low number of duplicate video sequences which suggests that the clustering approach is in this case suitable and valid.

On the other hand, the fraction of included information is rather low and suggests that higher speed-up factor or a function estimating significance of the video should be used. Further, the summaries contain relatively high amount of junk or information poor shots. This is probably also caused by the fact that the amount of significant information contained in the video is not currently estimated.

Table 1 Results achieved in the TRECVID 2008 [4] evaluations.

Meassure	Rank	Absolute value
Summary length	26	25.85
Judge time	26	40.56
Fraction of inclusions	15	0.40
Pleasant tempo	29	3.09
Duplicity	31	3.53
Junk	13	2.99

The table shows absolute values and relative ranks with respect to the results of other participants (higher rank is better; maximum rank is 43). The absolute values of last three measures are in the range 1-5 where 5 is the best.

6. CONCLUSION AND FUTURE WORK

This paper presents the solution for the rushes summarization task of TRECVID 2008, as it was developed by the Brno University of Technology team – Graph@FIT.

The system remained very simple – consisting basically of per-frame feature extraction and following clustering of groups of frames of constant time. From the previous year's version, mainly improvements of the clustering and feature extraction were incorporated, and the complex "passive user interface" layout was abandoned, or minimized to a thin timeline.

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 the shot's content distribution might help to choose the best summarization algorithm and evaluate the length of the final video. Setup parameters of summarization algorithms can be also tuned by analysis of the input 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.

The future work will also include better definition of the "variability" measure of the video shots that should better reflect the real variability of the content. Additionally, better definition of the cluster/shot boundaries and possibly also combination of clusters/shots will be added in order to avoid unnecessary subdivision of long shots.

A question raised from the observations of the final results, is 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 2008 organizers for taking the effort of preparing the evaluation and evaluating the results.

7. 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.

8. 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 Awad, G. The *TRECVid 2008 BBC rushes summarization evaluation.* In TVS '08: Proceedings of the International Workshop on > TRECVID Video Summarization, Vancouver, British Columbia, Canada, ACM, 2008, 1-20.

[5] Beran, V., Herout, A., Hradiš, M., Chmelař, P., Potúček, I., Sumec, S., Zemčík, P. Video Summarization at Brno University of Technology. *In: ACM Multimedia*, Augsburg, DE, ACM, 2007, 16-19.