This is a preliminary version of an article published in
Proc. of IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, June 2004
by Stephan Kopf, Thomas Haenselmann, Dirk Farin, Wolfgang Effelsberg

Automatic Generation of Video Summaries for Historical Films

Stephan Kopf, Thomas Haenselmann, Dirk Farin, Wolfgang Effelsberg
Praktische Informatik IV
University of Mannheim, 68131 Mannheim, Germany
{*kopf,haenselmann,farin,effelsberg*}@*informatik.uni-mannheim.de*

Abstract

A video summary is a sequence of video clips extracted from a longer video. Much shorter than the original, the summary preserves its essential messages. In the project ECHO *(European Chronicles On-line) a system was developed to store and manage large collections of historical films for the preservation of cultural heritage. At the University of Mannheim we have developed the video summarization component of the ECHO system. In this paper, we discuss the particular challenges the historical film material poses, and how we have designed new video processing algorithms and modified existing ones to cope with noisy black-and-white films.*

1 Introduction

In the project *ECHO* (European Chronicles On-line), a large software system was developed that stores and manages collections of historical films. The collections are precious from a cultural point of view since they document different aspects of life in European countries from the beginning of the last century until today. Video summaries support the users of the ECHO system.

Historical film archives are a great challenge for video analysis tools. Many well-known algorithms fail due to the properties of the old material (e.g., black-and-white films since they rely on color information). We have developed new algorithms that analyze such material reliably. Others than the ordinary features are required to find relevant shots in historical documentaries. A new heuristic approach is presented that selects the most important shots for the summary.

The remainder of this paper is organized as follows: Section 2 describes related work. We present the automatic analysis and selection of relevant shots in Sections 3 and 4, and conclude the paper with empirical results in Section 5.

2 Related Work

Many tools have been developed to generate a compact representation of a long video. The process is usually called *video summarization, video skimming* or *video abstracting*. The MoCA (movie content analysis) abstracting was one of the first tools to generate moving summaries from feature films automatically [4]. The system was initially developed to generate trailers of feature films. A major component was the detection of events of particular relevance such as explosions, gun-fire, or dialogs.

The Informedia Digital Video Library project at the Carnegie Mellon University has developed an interface to generate video skims [1]. Important words are identified in the textual transcript of the audio. Text and face recognition algorithms detect relevant frames which are combined into the final video skim. Sundaram et al. have presented algorithms for generating audio-visual skims [7]. Beside a robust audio segmentation, the visual complexity of a scene is analysed and a minimum time for comprehension is estimated.

Other projects developed algorithms to restore damaged or noisy films [3], but none of the existing research projects have addressed the specific challenges that arise from the analysis and summarization of historical films. Our experience shows that new algorithms must be developed and existing algorithms must be modified to cope with old films:

- most material is black-and-white, making color-based features useless,
- lots of noise is misleading the comparison of two adjacent frames,
- there is considerable jitter in the luminance. As a consequence many histogram-based techniques (e.g., for cut detection) fail.
- Films are often shaky, because hand-held cameras are used, making motion-based analysis much more difficult.
- Early camera men often made recording mistakes, e.g., the camera was pointed to the ground, and early film editors did not notice them or ignored them.
- Mistreatment in laboratories or early film projectors leads to scratches and stripes in the film.

3 Feature Extraction

The generation of a video summary is done in two steps: First, the video is analyzed and relevant features are calculated. In a second step, the most relevant shots are assembled to form the summary. Figure 1 gives an overview of

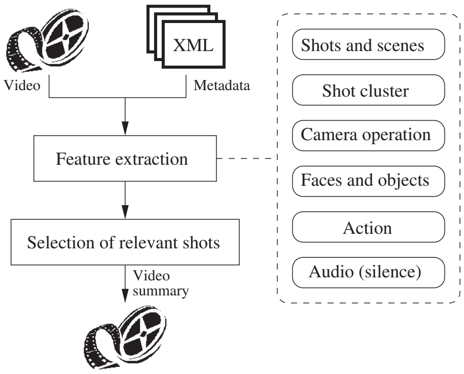

Fig. 1. Overview of the video summarization process.

this process. In this Section, we present algorithms to extract selected high-level semantic video features.

3.1 Camera Motion

Motion is one of the most important features in a film. We distinguish camera motion and object motion: if video is recorded with a moving camera, not only the objects in the foreground move, but those in the background as well.

We use a perspective camera model that has eight degrees of freedom to describe camera rotation (pan, tilt) or scaling (zoom-in, zoom-out). To calculate the camera parameters we identify a set of positions (features) in a frame that can be tracked throughout the shot. The Harris corner detector is employed to select appropriate feature points. Once the corners are identified, we establish correlations between corners in successive frames. In order to estimate camera parameters reliably from a mixture of background and object motion, we apply a robust regression method to estimate the eight parameters of the perspective camera model [2].

Our approach is very robust and handles historical films very well. Even in films with noise or jitter in luminance the camera motion can be estimated accurately. The approach fails if the film has scratches or stripes. Many corners are detected at the borders of these scratches and an exact estimation of the camera motion is not possible. We locate these faulty regions by temporal analysis of significant edges and remove the corners in these regions.

3.2 Recognition of Moving Objects

Our object recognition algorithm consists of two components, a segmentation module and a classification module. Figure 2 depicts the main recognition steps. The parameters of the camera motion are used to construct a background image for the entire sequence. The parameters are very precise and it is possible to construct a background image with shaky recordings of historical films. During construction of the background, foreground objects are removed by means of temporal filtering. Object segmentation is then performed by evaluating differences between the current frame and the constructed background image.

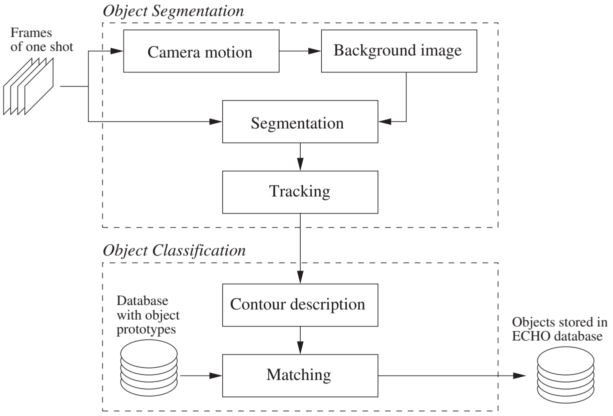

Fig. 2. Overview of the object recognition process.

Many frames in historical videos are noisy. On the other hand, the object segmentation algorithm is very sensitive to this noise since it is based on image differences. To reduce the effect of incorrectly detected object areas, a tracking algorithm that analyzes position, size, and geometry is applied to the object masks. Only objects that could be tracked through several frames of the shot are kept for further processing.

The classification module analyzes the segmented object masks. For each mask, an efficient shape-based representation is calculated (*contour description*) [5]. We use the curvature scale space (CSS) technique that is one technique to describe shapes in MPEG-7. It is based on the idea of curve evolution and provides a multi-scale representation of the curvature zero crossings of a closed planar contour. It is very robust to noise and works well with small or medium sized segmentation errors that are typical for historical films.

The matching process compares these contour descriptions to pre-calculated object descriptions stored in a database. The matching results for a number of consecutive frames are aggregated. This adds reliability to the approach since unrecognizable single object views occurring in the video are insignificant with respect to the entire sequence. A detailed description of the segmentation and classification algorithm can be found in [2, 5].

3.3 Shot Cluster

We define cluster as a syntactic grouping of frames based on a similarity measure, in contrast to a scene that is a semantic grouping.

The first problem is to identify appropriate key frames that represent a shot very well. Due to significant changes in luminance and single frame errors, the selection of the first or center frame of a shot is not possible. We select a frame as key frame that is most similar to the other frames in a shot. For each frame i, a feature vector H_i is extracted. We use quantized luminance histograms for 9 equal-sized regions as feature vector. The distance measure $D(H_i, H_j)$

is the sum of absolute differences. The key frame i is selected, that minimizes $\sum_j D(H_i, H_j)$.

We create a certain number of cluster centers. Each cluster center and key frame is represented as feature vector, that describes a position in a multi-dimensional space. The idea is to add new cluster centers till the distance of all key frames to the nearest center is very low.

We use a variation of the K-means algorithm to locate the cluster centers. The number of cluster centers is automatically estimated. The algorithm stops if the distances between key frames and the next cluster center are very low (all key frames in a cluster are very similar).

Many frames and shots in historical films are damaged. It is very important that these shots are not selected for the video summary. The clustering algorithm can be modified to identify them: Cluster centers are initialized with predefined shots that should *not* be part of the video summary; we call them *delete clusters.* Typical delete cluster centers are black, gray, or white frames. The selection process will discard them.

3.4 Face Detection

Persons are very important in most types of video, and especially in documentaries of historical value. Close-up views of the faces of main actors are important in feature films, whereas historical documentaries often feature sports persons, politicians, etc.

Rowley et al. [6] have developed a famous, very reliable face recognition algorithm based on a neural network. We have implemented the face detector and trained our own network with more than $7,500$ faces. The face detector does not rely on color information and works well with historical black-and-white videos.

4 Selection of Shots for the Summary

Figure 3 depicts the main steps of the selection process. In a first step, irrelevant shots are identified. Shots that have been attached to *deleted clusters* or very short shots (less than three seconds) are removed from the list.

We calculate aggregated feature values to make the different features comparable. Each value is normalized and characterizes a feature on the level of shots. Most aggregated feature values are initialized once and a modification is not required during the selection process (static features). Other feature values depend on previously selected shots (dynamic features). They are updated whenever a new shot is selected. In this section, we describe the idea of the selection process with few selected features. The summarization algorithm analyzes additional features like camera motion, action intensity, contrast, audio, etc.

4.1 Static Feature Values

The aggregated *face* value is the normalized quotient of face pixels to all pixels in a frame. With our definition, the relevance of two medium sized faces is similar to the relevance of one large face. The average value of all frames in a shot is stored as aggregated face value.

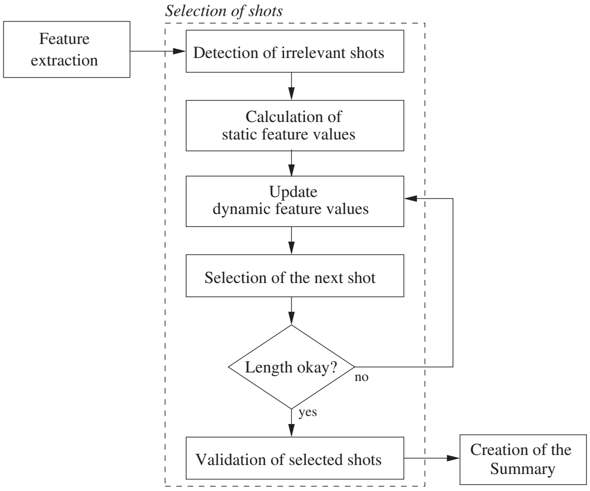
Selection of shots
Feature extraction
Detection of irrelevant shots
Calculation of static feature values
Update dynamic feature values
Selection of the next shot
Length okay?
no
yes
Validation of selected shots
Creation of the Summary

Fig. 3. Overview of the selection of shots.

Our *moving object* classification algorithm detects boats, planes, cars and people. These object classes were selected because they often appear in historical films (e. g., World War II). The aggregated value for moving objects is determined by the number of recognized objects in a shot, the size of the objects and the reliability of the recognition. If the recognition of an object is possible, we get additional information about the quality of a shot: A background image cannot be constructed with blurred frames (few corners are found in the background) and noise prevents an exact segmentation.

4.2 Dynamic Feature Values

The aggregated values, that have been described so far, are initialized once and the values do not change. Other feature values like shot clusters, scenes or position values are updated whenever a new shot has been selected. The relevance of a *cluster* C_i depends on the length of all shots that have been attached to cluster i:

$$C_i = \frac{L_i}{max\{L_j\}} \cdot \frac{1}{1+S_i}, \qquad j = 1 \ldots N,$$

where L_i is the summarized length of all shots of cluster i, S_i is the number of already selected shots from this cluster and N is the total number of clusters.

A major goal of a video summary is to give a general overview of the full video. Shots should be selected from all parts of the video. A summary of a feature film may have a different goal, because the thrilling end of a film should not be revealed. The $position$ value distributes the selected shots among the full length of the video. The value is a function that evaluates the distance from a shot to the next selected one.

4.3 Selection of Shots

The selection process uses the aggregated feature values $F_{i,j}$. The summarized relevance R_i for each shot i and feature j is defined as:

$$R_i = \sum_j \alpha_j \cdot F_{i,j}, \qquad \sum_j \alpha_j = 1,$$

α_j specifies the relevance of each feature for the summary. We have used fixed weights ($\alpha_j = \frac{1}{N}, j = 1 \ldots N$) in our implementation, although a user can modify them.

The selection algorithm is an iterative process as depicted in Figure 3. The static and dynamic feature values are calculated, and the shot with the maximum summarized relevance R_i is selected for the summary. The algorithm stops, if the summary has reached the desired length. Otherwise the dynamic feature values are updated and the next shot is selected.

4.4 Generation of the Summary

The last step defines the exact cut position (silent parts in the audio track are good candidates for a cut), selects the transitions between the shots, and creates the summary. Transitions in summary and film should be similar. The user can modify the frame-rate, resolution or bit-rate. E.g., if a user wants to create MPEG-I summaries with a lower resolution from high-resolution MPEG-II videos, he can specify the required parameters and the summary will be generated.

5 Experimental Results

Within the scope of the project *ECHO,* a system has been developed to store and manage large collections of historical films. Four major film archives[1] have selected very precious historical films. More than 100.000 hours of historical films are stored in these archives.

The ECHO system has stored more than 4500 films from 1920 to 1965 so far. For each new film, meta-data information is calculated and a summary is generated automatically. A user can generate his individual summary, e.g., a summary with all faces of the video.

The estimation of the camera operation is very precise. Errors only occur in case of large moving foreground objects or blurred background images.

Our face detection system locates about 90 percent of the faces with a width and height of at least 25 pixels. The recognition rate of moving objects depends on the object and shot: It is acceptable in shots with one car or one person (about 40 percent), but much lower for planes or boats due to difficult background (water) or missing edges (e.g., sky with some clouds). Many objects were missed, but nearly no wrong classification occurs.

We have received feedback from other partners of the ECHO project and made some local tests. Additionally, an extensive test of the system was performed with 17 professional users (5 cataloguers, 12 editors). In general, the quality of the summaries is very good and the essential message of the original video is preserved. Several cataloguers noticed that the summaries may be useful to make a textual description of the video. On the other hand, some editors reported that the risk of missing (possibly relevant) parts of the video is too big. They do not trust automatically generated video summaries and prefer to work on the original material. In some cases, important parts of the film were missing and the understanding of the content of the summaries was very difficult. This is a typical problem of very short summaries.

The selection of shots is very subjective and an optimal summary is not possible. Two persons will select different shots from a long documentary, because they rate their importance differently. An automatic generated summary makes a third – not necessarily optimal – selection.

Acknowledgments

The work of Stephan Kopf was financially supported by the European project ECHO. We thank the colleagues in the ECHO project – especially from Instituto Luce, Memoriav, NAA and INA – for the film material.

6 References

[1] Michael G. Christel, Alexander G. Hauptmann, Adrienne S. Warmack, and Scott A. Crosby. Adjustable filmstrips and skims as abstractions for a digital video library. In *Proceedings of the IEEE Advances in Digital Libraries Conference*, pages 98–104, 1999.

[2] Dirk Farin, Thomas Haenselmann, Stephan Kopf, Gerald Kühne, and Wolfgang Effelsberg. Segmentation and classification of moving video objects. *Handbook of Video Databases*, pages 561–591, 2003.

[3] Anil C. Kokaram, Rozenn Dahyot, Francois Pitie, and Hugh Denman. Simultaneous luminance and position stabilization for film and video. In *Visual Communications and Image Processing*, January 2003.

[4] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. In *Communications of the ACM*, volume 40, pages 55–62, 1997.

[5] Stephan Richter, Gerald Kühne, and Oliver Schuster. Contour-based classification of video objects. In *Proceedings of SPIE, Storage and Retrieval for Media Databases*, volume 4315, pages 608–618, January 2001.

[6] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 20, pages 23–38, 1998.

[7] Hari Sundaram, Lexing Xie, and Shih-Fu Chang. A utility framework for the automatic generation of audio-visual skims. In *Proc. 10th SIG ACM Conference On Multimedia (ACM Multimedia 2002)*, Dec. 2002.

[1] Instituto Luce (Italy), Memoriav (Switzerland), Netherlands Institute for Sound and Vision (the Netherlands), and Institut Nationale de l'Audiovisuel (France).