# Analysis of Historical Artistic Documentaries

Matthias Zeppelzauer, Dalibor Mitrovic, and Christian Breiteneder
Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstrasse 9-11, A-1040 Vienna, Austria
{zeppelzauer, mitrovic, breiteneder}@ims.tuwien.ac.at

## Abstract

*The paper introduces a novel interdisciplinary project addressing the analysis of historical artistic films. The type of employed material has not been subject to automatic analyses, so far. It poses challenges in all areas of content-based analysis and retrieval due to its complex temporal structure and due to substantial degradations. We propose robust features and a method for shot cut detection for this material that outperforms established techniques.*

## 1. Introduction

In this paper, we introduce the project *Digital Formalism: The Vienna Vertov Collection*[1]. The project is a joint effort of film scientists, archivists and computer scientists with the goal of gaining insights into the work of the Russian avant-garde filmmaker Dziga Vertov (1896-1954). Vertov is famous for his highly formalized style of making films containing spatio-temporal structures and montage patterns that follow complex (sometimes hidden) rules and artistic principles. In the context of this project, film scientists provide the tasks and requirements for the automatic analysis.

From the perspective of computer science the project offers a novel type of material, expert annotations and real-world problems stated by film scientists. The films differ in structure and quality from the material traditionally employed in content-based retrieval, such as TRECVID material or Hollywood film, news broadcasts, and sports videos. Vertov's films are historical artistic documentaries from the 1920s and 1930s. Historical film in general has rarely been the subject of automatic content analysis [5, 8].

In this paper, we focus on the detection of shot cuts as one of the first tasks in the context of this project. The detection of shot boundaries is the basis for most high-level investigations, such as analysis of shot rhythm, montage patterns, and duplicate shot detection.

Most state-of-the-art shot boundary detection algorithms incorporate color information, e.g. 15 out of 19 shot boundary detection algorithms in TRECVID 2006 and 8 out of 11 in TRECVID 2007. Since the subject of this project is black and white material it is not safe to assume that the color-based shot boundary detection techniques are applicable. Therefore, we propose a technique for shot cut detection that does not rely on color and that is robust against the various artifacts in the addressed material. The technique outperforms established shot cut detection algorithms.

The contributions of this paper are the presentation of a novel project and challenging material to the scientific community as well as the development of a method for shot cut detection that is able to cope with the issues related to historical, archive films.

The remainder of this paper is organized as follows. Section 2 discusses the novel material in detail. In Section 3 we present the proposed method for shot cut detection. We discuss experimental results in Section 4 and give related work in Section 5. Finally, we draw conclusions and propose future work in Section 6.

## 2. Material

The films in this project are artistic documentaries produced in the Soviet Union in the 1920s and 1930s. In contrast to news- and sports broadcasts and feature film, these documentaries do not contain any narrative structure. In fact the director Dziga Vertov strongly opposed any narration in the Hollywood sense. Consequently, the films share a collage-like style.

The original material is 35mm black-and-white film to be played at non-standard frame rates of 18 to 21 frames per second. Most of the films are silent. We scan the filmstrips frame-by-frame in order to obtain an image sequence

[1] http://www.digitalformalism.org. Project partners: Austrian Film Museum, University of Vienna, Department for Theatre, Film and Media Studies, and Vienna University of Technology, Interactive Media Systems Group.

978-0-7695-3130-4/08 $25.00 © 2008 IEEE
DOI 10.1109/WIAMIS.2008.11

for each film. This process avoids the introduction of interpolated frames which usually occur during digitization at standard frame rates (e.g. 25fps for PAL). Interpolated frames represent information that does not exist in the original material. It is crucial to note that Vertov employed the number of frames for a given shot as a stylistic device. Thus interpolated frames would tamper with the intended statements of the films.

The filmstrips are multiple-generation copies that were never intended to be used for other purposes than backups. Due to this fact, these copies were not handled with much care in the film archives. Today, the original films do not exist any more, hence the available backup copies are the only existing source material left. The state of the material has degraded significantly, during storage, copying, and playback over the last decades (see Figure 1). An important issue in that context is the shrinking of the organic film. Shrinking refers to the process of physical contraction (horizontally and vertically) of the filmstrip over time which introduces frame displacements and non-linear geometric distortions. The most common artifacts we have to deal with, are:

- Scratches, introduced by dirt in the projectors during playback.
- Framelines visible in the frames due to copying misaligned filmstrips.
- Oversaturated frames due to degradation of contrast during copying.
- Dirt (dust, liquids, etc.) copied into the images.
- Frame displacements introduced by shrinking and copying under suboptimal conditions. Note the varying position of the framelines in Figure 1.
- Flicker, inherent in the old material.

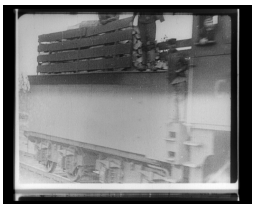
All these artifacts have been accumulated with every generation of copies. They considerably influence shot boundary detection. Dirt, dust, liquids (spilled over the filmstrips), and scratches introduce noise that generates abrupt visual changes. These unintended changes interfere with established shot cut detection algorithms that are based on pixel differences, edges, and corners (feature points). Additionally, frame displacements disturb techniques that rely on motion information. Flicker globally influences the distribution of intensity values, which limits the power of histogram-based approaches. Kopf et al. discuss several other issues in the context of the analysis of old films [5].

Another issue related to this material are its complex montage patterns. Vertov's films regularly contain sequences of short shots (one to four frames). This introduces problems for shot cut detection techniques that rely on large processing windows.

From the state of the material, we draw the conclusion that there is a demand for adaptation of existing and development of novel techniques for film analysis. We aim at developing a shot cut detection method that is robust against the mentioned local and global artifacts and is able to cope with the complex temporal structure of the material.

(a) dirt (top left), vertical scratch (right), frameline copied into the image (bottom)


(b) tear in the middle of the frame



(c) artifact introduced during copying (error in brightness)

(d) unwanted high contrast, visible frameline (bottom)

**Figure 1. Artifacts often found in archive film.**

## 3. Proposed Method

The proposed method is based on a technique originally developed for music information retrieval by Foote [4]. Foote detects segment boundaries in music by computing and analyzing self-similarity in the audio spectrum. We adapt this technique for shot cut detection for highly degraded black-and-white film.

### 3.1. Feature Extraction

We propose two features for shot cut detection that are robust against the artifacts present in the material. The low-frequency content of the frames is captured by a discrete cosine transform (DCT) feature, while the high-frequency content is represented by an edge descriptor.

For the DCT feature, we uniformly split each frame into $B$ image blocks. We transform each block into frequency domain by a DCT and extract the first $N$ low-frequency coefficients. The coefficients of all blocks yield a $B * N$-dimensional feature vector.

The parameter $N$ should be chosen in a way that high-frequency distortions are removed. Parameter $B$ determines the block size. It balances the influence of frame displace-

ments and motion in a block and the amount of preserved spatial information. Large blocks lead to high robustness against frame displacements, but a loss of spatial information while small blocks lead to the opposite.

The DCT feature represents the coarse intensity distribution among the blocks of the frames. It is robust against local high-frequency artifacts, such as dirt and scratches. Furthermore, it compensates for frame displacements and flicker to a high degree.

The second feature captures the orientations of the edges in the frames. Edges contain salient information for shot cut detection. They represent semantically meaningful information, such as contours and object boundaries that usually change considerably across shot cuts [6]. We employ an edge histogram, similar to the MPEG-7 Edge Histogram. The edge histogram (EH) is computed for the same $B$ blocks as the DCT feature. The histogram of each block contains 5 bins, for horizontal, vertical, 45 degree, 135 degree, and non-directional edges. The edge histogram for the entire frame contains $B * 5$ bins.

The EH represents the distribution of orientations of the edges across the blocks of the frame. It is highly robust to frame displacements since it captures global information within in each block. Additionally, the feature is invariant to flicker. The EH captures high-frequency information which makes it prone to artifacts like scratches and dirt. The influence of these artifacts is usually low compared to the influence of the dominant and meaningful edges. However, global artifacts, such as tears across the entire frame are reflected in the feature (see Figure 1(b)).

The DCT feature and the EH are well suited for combination, since they capture orthogonal and complementary information. The DCT feature represents low-frequency information, while the EH summarizes high-frequency content.

### 3.2. Similarity Comparison

We compute the similarity between feature vectors of adjacent frames similarly to Cooper and Foote [2]. First extract a one dimensional feature vector (e.g. EH, DCT, or a combination of both features) for each frame. Then the pairwise similarity of all feature vectors is computed by the cosine measure. This results in a (symmetric) similarity matrix with maximum values at the diagonal (see Figure 2(a)). Each entry in the matrix corresponds to the similarity of two feature vectors of two frames. Time progresses along the rows and the columns of the matrix, as well as along the main diagonal. Similar frames yield high values (high similarity) while dissimilar frames yield low values (low similarity) in the matrix. Sequences of similar frames (e.g. frames of a shot) produce bright squares along the diagonal. This results in a checkerboard pattern, as shown in Figure 2(a) where the bright squares correspond to shots of the film. We detect shot cuts by moving a square, Gaussian weighted checkerboard kernel of size $W$ along the diagonal of the similarity matrix, illustrated in Figure 2(b). The checkerboard kernel yields high correlation at the shot cuts and low correlation at other positions. The correlation of the checkerboard kernel is a one-dimensional function $C$ over all frames.

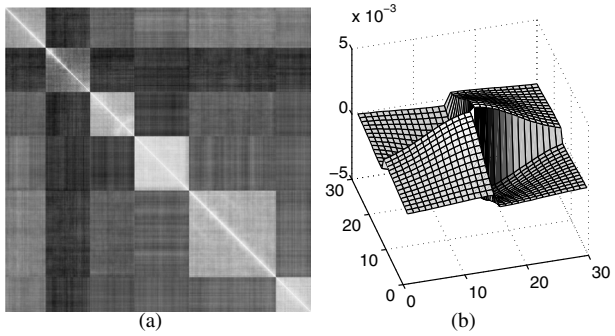



**Figure 2. (a) shows an excerpt of the similarity matrix of The Eleventh Year. The bright squares indicate shots. In (b) a 30x30 Gaussian weighted checkerboard kernel is illustrated**

Computation of the entire similarity matrix is much too expensive and would require excessive amounts of memory. In practice, it is sufficient to compute similarities near the diagonal of the similarity matrix. The size $W$ of the checkerboard kernel, defines the size of the analysis window. Further details regarding the construction and computation of the similarity matrix and the kernel correlation are given in [2].

### 3.3. Shot cut detection

The correlation of the checkerboard kernel is a one-dimensional function $C$ over all frames. The function shows peaks at potential shot cuts and has values near zero in homogeneous areas. A point in the correlation function is considered a shot cut if it is a local maximum and the difference to the preceding value exceeds a threshold $t_c$. The peak detection results in the list of detected shot cuts.

### 3.4. Feature Combination

We propose the combination of the two features for shot cut detection. There are several ways to merge the information contained in both features. One way, is to merge both features into one vector and then perform similarity comparison and shot cut detection. However, similarity computation of two feature vectors reduces the information con-

tained in the vectors to a single value. Consequently information captured by the features is lost at an early stage of processing.

Another possibility is to perform similarity comparison in parallel and independently for both features and to merge the resulting kernel correlation functions afterwards. Therefore, we compute two similarity matrices and filter them separately with the checkerboard kernel. This results in two kernel correlations, $C_{DCT}$ and $C_{EH}$ for the DCT feature and the edge histogram, respectively. We merge both correlation functions by a linear combination:

$$C_{merged} = w_{DCT} * C_{DCT} + w_{EH} * C_{EH},$$

where $w_{DCT}$ and $w_{EH}$ are the respective weights. The merged correlation function $C_{merged}$ is employed for shot cut detection as discussed in Section 3.3. The advantage of this approach is that more feature information is preserved for the actual shot cut detection in Section 3.3.

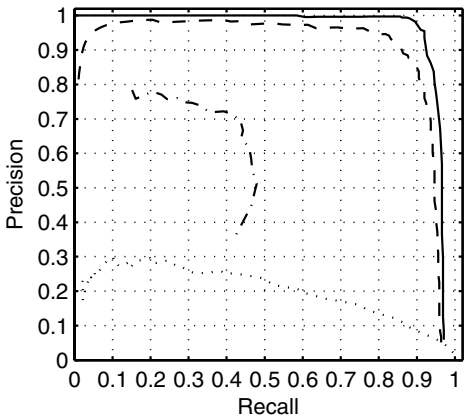
## 4. Experimental Results

We compare the proposed method with established, readily available techniques, namely a feature-based algorithm (Edge Change Ratio - ECR) proposed by Zabih et al. [11], the MoCA shot cut detector (employs histograms) [7] and a block-based histogram technique [1].

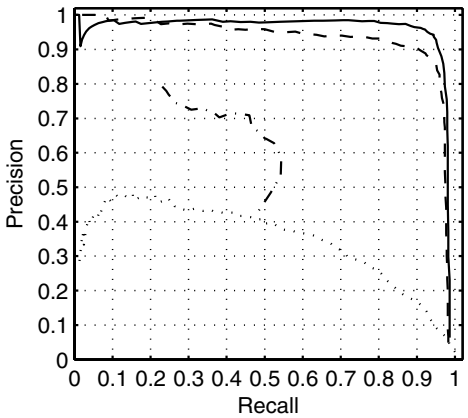
The parameters for the proposed technique are chosen as follows: The number of image blocks $B$ determines the robustness to frame displacements and motion. A value of $B = 9$ has shown to be a good tradeoff. The size of the kernel $W$ has to be proportional to the length of the shortest shots in the films. A small kernel is necessary in order to detect short shots of only a few frames. A kernel size of $W = 6$ frames provides a good tradeoff between detection rate and noise robustness. A number of $N = 36$ low-frequency DCT coefficients is suitable for the material employed in the experiments. The weights $w_{DCT}$ and $w_{EH}$ of the linear combination are chosen to be 0.5 (variations did not result in improved performance).

Two films have been digitized and annotated by domain experts in the course of the project so far. We evaluate the shot cut detection techniques for Vertov's films *Kinopravda 21* and *The Eleventh Year*. Kinopravda 21 has 35060 frames ($\approx$ 32min at 18fps) and contains 411 hard cuts, The Eleventh Year is 63123 frames long ($\approx$ 58min at 18fps) and contains 646 hard cuts.

We apply the proposed method to both films and compare the results with that of the above mentioned techniques. Performance is measured in terms of recall and precision. We build recall-precision graphs by varying the threshold $t_c$ (see Section 3.3). Figures 3(a) and 3(b) show recall versus precision of the employed techniques for both films.



(a) Kinopravda 21



(b) The eleventh year

**Figure 3. Recall-precision graphs for both films. The solid line is the proposed method, the dashed line is ECR, the dotted line is the histogram-based approach and MoCA is the dash-dot line.**

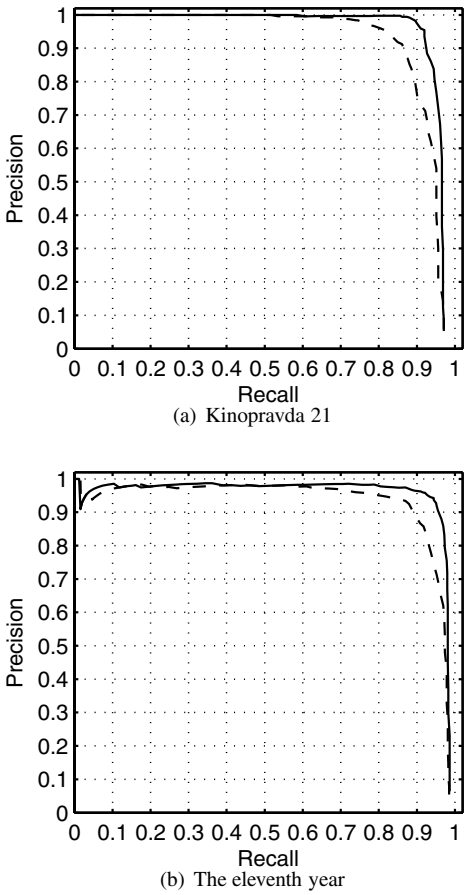
We observe that the histogram-based techniques are not appropriate for shot cut detection in this material. The main reason for this is that the histograms are not robust against flicker. ECR yields significantly higher recall and precision than the histogram-based techniques. This proves the assumption that edge information is more robust to the artifacts in the films. However, the proposed technique outperforms ECR in recall as well as in precision for both films. For the film Kinopravda 21 both measures are significantly increased compared to ECR. In the film The Eleventh Year, the proposed method mainly increases precision. These results are promising in the context of the highly degraded material. We summarize the performance of the discussed methods in Table 1. We compute the F1-scores for all recall and precision pairs obtained in the experiments and list the maximum F1-scores, in order to provide a measure of the achievable performance.

**Table 1. The maximum F1-scores obtained from the recall-precision pairs for all investigated methods**

| Method | Kinop. 21 | 11th. Year |
|---|---|---|
| Histogram-based | 0.32 | 0.46 |
| MoCA | 0.54 | 0.57 |
| ECR | 0.88 | 0.91 |
| EH | 0.86 | 0.86 |
| DCT | 0.89 | 0.89 |
| EH+DCT (early merged) | 0.89 | 0.90 |
| EH+DCT (lin. comb.) | **0.94** | **0.94** |

We further analyze the performance of some variations of the proposed method. As mentioned in Section 3.1 the DCT feature and the EH describe orthogonal information and thus are candidates for combination. We compare the performance of the shot cut detector based on the isolated features with that of the combination. Experiments show that both features separately yield only suboptimal results. The linear combination of the kernel correlations raises the performance figures significantly, which proves its beneficial effect. Table 1 lists the respective performance figures in rows *EH*, *DCT*, and *EH+DCT (lin. comb)*.

We have identified two strategies for the combination of the features in Section 3.4. The performance of both approaches is evaluated. The combination of the features prior to similarity comparison (*EH+DCT (early merged)*) yields only suboptimal results. As expected, the second strategy (linear combination of individual kernel correlations) significantly increases the performance. The corresponding recall-precision graphs are depicted in Figure 4.



(a) Kinopravda 21

(b) The eleventh year

**Figure 4. Recall-precision graphs for both films. The solid line is the proposed method with linear combination of the kernel correlations, the dashed line is the proposed method with merging of the features prior to similarity comparison.**

## 5. Related Work

Shot boundary detection is a well-investigated topic in video analysis. Extensive work has been conducted so far [1, 10]. Most state-of-the-art shot cut detectors rely on color information which may be observed from the submissions to the last TRECVID shot boundary detection tasks. Only a small number of these techniques supports black-and-white material. Generally, even fewer techniques target the special case of archive film. Urhan et al. propose novel techniques for shot cut detection geared towards old film [8, 9]. Their approaches exploit phase correlation and kernel-based comparison to detect hard cuts in visually degraded and distorted films. Kopf et al. perform summarization of historical archive films based on shots [5].

Cooper and Foote propose a shot boundary detection technique based on prior work in music information retrieval [2, 3, 4] for color videos. The authors compute the similarity of adjacent frames and construct a similarity matrix. They analyze the similarity matrix with different kernels and machine learning methods in order to detect shot boundaries.

Low-level features are the basis for the construction of the similarity matrix. Cooper and Foote mainly propose color and grayscale histogram features for this purpose. These features are not applicable to the employed material, where flicker and intensity variations are omnipresent. Another feature proposed in [2] are the global low-order discrete cosine transform (DCT) coefficients of the three color channels. This feature is not appropriate in black-and-white film where only one channel exists. The global DCT coefficients of this single channel do not capture enough discriminatory information for shot cut detection.

## 6. Conclusions and Future Work

We have presented a novel project in the domain of content-based video analysis, that focuses on historical artistic documentaries by Dziga Vertov. The project provides novel material that poses challenges for content-based analysis. We have presented first promising results for shot cut detection. We propose novel features and a technique that is suitable for the complex spatio-temporal structure and the manifold artifacts of the films.

In the future, we will compare the proposed method with state-of-the-art shot cut detection techniques and investigate the influence of missing color information and the effects of the described artifacts on the detection rates. Additionally, we will evaluate the proposed method with different types of material in order to show its general applicability.

## Acknowledgments

This work has received financial support from the Vienna Science and Technology Fund (WWTF) under grant no. CI06 024.

## References

[1] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, 5(2):122–128, 1996.

[2] M. Cooper and J. Foote. Scene boundary detection via video self-similarity analysis. In *Proc. of the Int. Conf. on Image Processing*, volume 3, pages 378–3813, 2001.

[3] M. Cooper and J. Foote. Video segmentation via temporal pattern classification. *IEEE Trans. on Multimedia*, 9(3):610–618, 2007.

[4] J. Foote. Visualizing music and audio using self-similarity. In *Proc. of the 7th ACM Int. Conf. on Multimedia*, pages 77–80. ACM, 1999.

[5] S. Kopf, T. Haenselmann, D. Farin, and W. Effelsberg. Automatic generation of video summaries for historical films. In *Proc. of the Int. Conf. on Multimedia and Expo*, volume 3, pages 2067–2070, 2004.

[6] R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. *Int. Journal of Image and Graphics*, 1(3):469–486, 2001.

[7] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Commun. ACM*, 40(12):54–62, 1997.

[8] O. Urhan, K. M. Gullu, and S. Erturk. Modified phase-correlation based robust hard-cut detection with application to archive film. *IEEE Trans. on Circuits and Systems for Video Technology*, 16(6):753–770, 2006.

[9] O. Urhan, K. M. Gullu, and S. Erturk. Shot-cut detection for b&w archive films using best-fitting kernel. *Int. Journal of Electronics and Communications*, 61(7):463–468, 2007.

[10] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A formal study of shot boundary detection. *IEEE Trans. on Circuits and Systems for Video Technology*, 17(2):168–186, 2007.

[11] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *Proc. of the 3rd ACM Int. Conf. on Multimedia*, pages 189–200, New York, NY, USA, 1995. ACM.