Gradual Transition Detection in Historic Film Material—A Systematic Study

MARKUS SEIDL, University of Applied Sciences St. Pölten
MATTHIAS ZEPPELZAUER, DALIBOR MITROVIĆ, and CHRISTIAN BREITENEDER,
Vienna University of Technology

The segmentation of films and videos into shots requires the detection of gradual transitions such as dissolves and fades. There are two types of approaches: unified approaches, that is, one detector for all gradual transition types, and approaches that use specialized detectors for each gradual transition type. We present an overview on existing methods and extend an existing unified approach for the detection of gradual transitions in historic material. In an experimental study, we evaluate the proposed approach on complex and low-quality historic material as well as on contemporary material from the TRECVid evaluation. Additionally, we investigate different features, feature combinations, and fusion strategies. We observe that the historic material requires the use of texture features, in contrast to the contemporary material that, in most of the cases, requires the use of color and luminance features.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Shot boundary detection, gradual transition detection, cultural heritage

ACM Reference Format:

Seidl, M., Zeppelzauer, M., Mitrović, D., and Breiteneder, C. 2011. Gradual transition detection in historic film material—A systematic study. ACM J. Comput. Cult. Herit. 4, 3, Article 10 (December 2011), 18 pages.
DOI = 10.1145/2069276.2069279 http://doi.acm.org/10.1145/2069276.2069279

1. INTRODUCTION

Films have become an integral part of our cultural heritage. Today, vast amounts of (black and white) films are stored in specialized archives and museums. These archives and museums have begun to digitize their collections in an effort to enable a broader access to the films. However, digitized content requires additional postprocessing to make it truly accessible. For large amounts of digitized data, this postprocessing needs to be as automated as possible because human resources are expensive and limited. For example, to enable nonlinear browsing of digitized film material, the material needs to be segmented into smaller units, such as shots. In this article, we present and evaluate automated gradual transition detection for the segmentation of digitized historic films.

Contact author's e-mail address: markus.seidl@fhstp.ac.at.

DOI 10.1145/2069276.2069279 http://doi.acm.org/10.1145/2069276.2069279

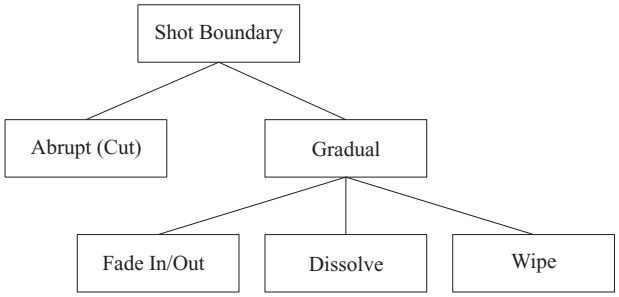

Fig. 1. A taxonomy of shot boundaries. Shot boundaries between shots are either abrupt transitions (called cuts) or gradual transitions. We distinguish between three structurally different types of gradual transitions: fades (either fade in or fade out), dissolves, and wipes.

A shot is defined as a continuous sequence of frames which have been captured in one camera action. The transitions between shots can either be abrupt or gradual (see Figure 1), and are referred to as either *cut* or *gradual transition*.

The National Institute of Standards and Technology (NIST) started the TRECVid benchmark for content-based retrieval in 2001 [Smeaton et al. 2006]. Since then, a large amount of test footage for shot boundary detection became available, and enabled the objective comparison of different approaches. Since 2005, the detection of cuts can be considered as solved [Yuan et al. 2005]; since 2008, the detection of gradual transitions is also declared solved [Smeaton et al. 2006, 2010].

For historic material, we agree in the case of cuts. We achieved satisfactory results in previous work on cut detection in historic material [Zeppelzauer et al. 2008]. However, we disagree for the case of gradual transition detection in historic material. Shot boundary detection for historic material presents novel challenges. This type of material has special properties that we have to take into account. For example, artifacts like scratches, mold, flicker, and shaking as well as complex and long gradual transitions. We might expect that a preceding film restoration solves most of these problems. However, film restoration is a time-consuming and expensive process which usually requires human interaction. We aim at the development of fully automatic methods that enable efficient analysis of large amounts of film material (entire archives). Consequently, restoration is not feasible in this scenario. In this article we present a broad experimental study that shows how the characteristics of historic material influence the steps of the gradual transition detection process.

This article is organized as follows. In Section 2 we define the problem of gradual transition detection. Section 3 presents the state-of-the-art in shot boundary detection. In Section 4, we discuss the method for gradual transition detection in historic material. Section 5 describes the experimental study. In Section 6, we present the most important results, and draw conclusions in Section 7.

2. PROBLEM DESCRIPTION AND MATERIAL

Gradual transition detection for contemporary video material is a well-investigated problem. We briefly state the problems of gradual transition detection in general and explain the specific problems in the context of historic material.

2.1 Problem Description

The main problems in gradual transition detection are (i) many different gradual transition types exist; (ii) the gradual transitions have varying lengths; and (iii) object- and camera movements are easily confused with gradual transitions [Yuan et al. 2005; Hanjalic 2002; Lienhart 2001]. These three main

Table I. Gradual Transition (GT) Types and their Lengths in Historic and Contemporary Material

Material	GT Type	#Frames / GT		
		Min	Max	Mean
Historic	Dissolve	15	134	31.8
	Iris In (From Black)	10	52	26.3
	Complex Transition	25	93	49.6
	Iris In (Not Black)	23	34	30.9
	Bar Wipe	25	53	36.8
	Iris Out (To Black)	6	21	12.5
	Iris Out (Not Black)	29	47	38.0
	Fade Out	20	25	22.5
	All	*6*	*134*	*30.7*
TRECVid	Dissolve	1	22	2.9
	Fade in/out	7	16	10.9
	Other	4	107	21.7
	All	*1*	*107*	*11.83*

The gradual transition types are sorted by descending frequency of occurrence for historic and TRECVid material, respectively.

problems are described well by Yuan et al. [2007]. Additionally to these three problems, two specific problems occur with historic material: first, the historic material includes longer and differing types of gradual transitions compared to the contemporary one, and second, the historic material contains many artifacts that interfere with gradual transition detection. Table I shows the lengths of different gradual transitions in contemporary and historic material. We observe that the mean lengths for historical material are larger than for contemporary material.

2.2 Material

The historic material originates from the black and white film *Kinoglaz* [Vertov 1924] by the Soviet filmmaker Dziga Vertov. It was produced in 1924 and has a runtime of 78 minutes at 18fps. Due to the film's black and white nature, content-based features that rely on color information are not applicable. The copy is several decades old and contains the following artifacts.

—Flicker. Due to manual film transport in old cameras, the exposure along the filmstrip is unsteady; see Figure 2.

—Image vibrations. The shrinking of the film over the last decades—due to chemical properties of the filmstrips, the images are shaky; see Figure 3.

—Degraded contrast. Due to the fatigue of the material, the contrast of films is degraded.

—Scratches. The films have long vertical scratches due to mechanical problems in the old projectors; see Figure 4(a).

—Mold/dirt. The films contain dirt and a visible mold is growing on the material; see Figure 4(b).

—Wrong exposure or development. Unreliable light meters and unreliable camera shutters or nonstandardized development under nonstable conditions caused wrong exposure or development.

3. RELATED WORK

We survey and compare two principally different approaches: specialized approaches, that is, one detector for each gradual transition type, and unified approaches, that is, one detector for all gradual transition types. We focus on specialized approaches for dissolve detection, since dissolves are the most common gradual transitions.

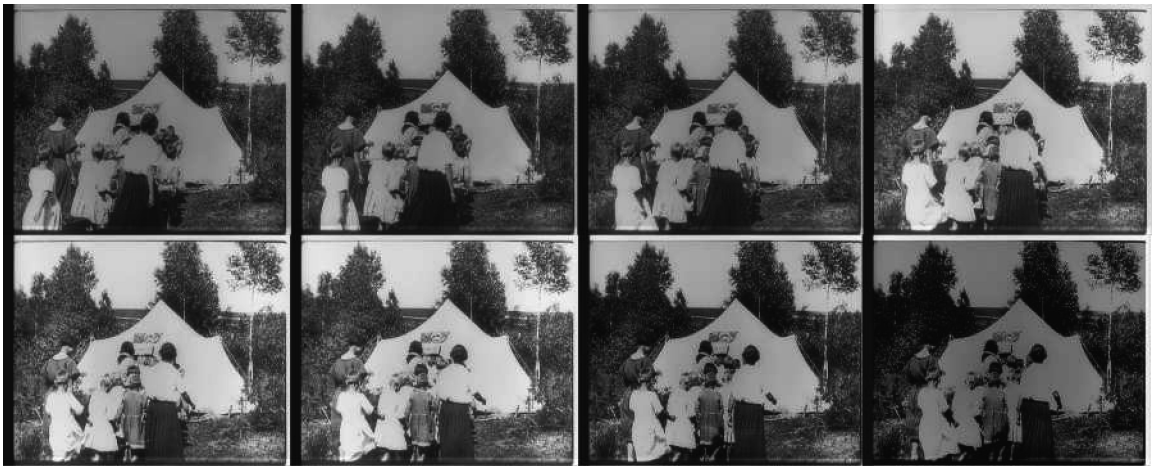

Fig. 2. This sequence of frames illustrates flicker in the historic material. The brightness of subsequent frames changes rapidly due to degradation. Kinoglaz [Vertov 1924] frames reproduced by courtesy of Österreichisches Filmmuseum Wien/Austrian Film Museum Vienna.

Fig. 3. Two subsequent frames from the film *Kinoglaz's* beginning titles. The white horizontal line serves as a reference for the comparison of the left and right frames. Notice that the text in the frame to the right is shifted downwards in comparison to the frame to the left. Kinoglaz [Vertov 1924] frames reproduced by courtesy of Österreichisches Filmmuseum Wien/Austrian Film Museum Vienna.

The twin comparison method for the detection of dissolves [Zhang et al. 1993] was one of the first gradual transition detection methods. It utilizes accumulated interframe differences to achieve intertemporal comparison longer than one frame. It uses color histograms as features. The detection is done by thresholding.

Yuan et al. improve the twin comparison method for gradual transitions lasting longer than five frames [Yuan et al. 2004]. The approach includes motion vectors that control a self-adapting threshold. The features used are color histograms, difference images, and motion vectors.

Kawai et al. [2007] propose a method that is based on two assumptions. First, if a shot changes to another shot, the luminance of each pixel increases or decreases monotonically. Second, the comparison of a dissolve with an ideal dissolve yields a small error rate. The features used are RGB pixel values and RGB histograms. The detection is done with suitable thresholds. The postprocessing steps are heuristical verification (duration of a dissolve) and a comparison of the similarity of the begin and end frames of the dissolve candidate.

(a) (b)

Fig. 4. In the left frame two significant vertical scratches are highlighted by black ellipses. The dark scratchmark is introduced by copying from a scratched source. The right frame contains several artifacts introduced by dirt which are highlighted by white ellipses. Kinoglaz [Vertov 1924] frames reproduced by courtesy of Österreichisches Filmmuseum Wien/Austrian Film Museum Vienna.

The three approaches mentioned so far are based on color information and employ thresholds for detection. The usage of thresholds causes a lack of robustness. Yuan et al. [2005] state that a threshold's value depends highly on the genre of the video and that a threshold cannot make use of the information if a valley or peak is sharp or gentle.

Liu et al. delivered the best shot boundary detection performance in TRECVid 2006 and 2007. They propose a dissolve detector based on the change of the color histogram variance during a dissolve [Liu et al. 2006, 2007]. They assume that a dissolve is a linear mixture of two shots and hence that the change of the color variance during a dissolve follows typical curves. These curves are detected by finite state machines. The detector extracts color histograms and edge histograms of each frame, and calculates motion compensated intensity matching errors and histogram changes between the current frame and its first predecessor as well as the current frame and its sixth predecessor. The classification is done with finite state machines and support vector machines (SVM).

Bescos et al. [2005] introduce a *unified* approach for gradual transition detection which is based on the patterns that result from interframe comparison with different temporal distances. They use RGB color values as features and classify by thresholding.

Yuan et al. [2005] utilize a similarity matrix with interframe similarity. The approach is based on the fact that due to the varying length of gradual transitions, a gradual transition does not leave a pattern as clear as a cut in the similarity matrix. Therefore a self-similarity matrix is calculated in lower resolution, for example, by decreasing the video sampling rate. In this low-resolution similarity matrix a gradual transition leaves a clearer pattern. The features employed are global and block-based color histograms; the classification is done with an SVM.

Cooper et al. [2007] propose a well-performing shot boundary detection approach that calculates a self-similarity matrix containing the interframe similarities of all frames of a sequence. The detector uses the area in the similarity matrix surrounding the frame as an intermediate feature. The intermediate features are used for classification of the frames with k-NN. The features for the calculation of the similarity matrix are global and block-based color histograms. The verification step employs temporal heuristics.

Specialized approaches are more complex, as they use a single detector for each gradual transition type and have to merge results in an additional step. Each gradual transition type requires a special detector, and each detector needs training. The unified approaches consist of fewer processing steps than the specialized ones. They are more general, and therefore more transparent. As the unified

approaches rely on intertemporal comparison of more than two frames, we expect them to be more robust against distorted material. The specialized approaches have a better detection performance. As unified and specialized approaches rely mainly on color values and histograms as features, we expect both to be equally incompatible to black and white material.

The approach of Bescos et al. [2005] relies on intertemporal comparison of all frames with *one* frame. We expect a lack of robustness for distorted material. The approaches of Yuan et al. [2005] and Cooper et al. [2007] rely on self-similarity matrices that compare all frames of a sequence with each other. Both approaches are tested against the TRECVid material and perform comparably well. We expect these two approaches to be superior to the approach of Bescos et al. [2005] for distorted material.

We adapted the approach of Cooper et al. [2007] to the detection of abrupt transitions in historic material in Zeppelzauer et al. [2008]. The approach yields satisfactory results for cut detection. In this article, we extend our work to the detection of gradual transitions in historic material.

4. PROPOSED METHOD

According to Yuan et al. [2007] a shot boundary detector consists of three processing steps: (i) visual content representation; (ii) construction of the continuity signal; and (iii) classification. We add a fourth step (iv) of verification, as we perform a postprocessing step for verification of the classification results.

4.1 Visual Content Representation

Feature extraction aims at representing the visual content of the images in a compact yet informative way. The special requirements of our approach are: (i) invariance towards flickering frames, scratches, mold and dust; (ii) some features should be invariant to object motion and camera motion; (iii) some features should be sensitive to object motion and camera motion. The second and third requirements seem contradictory. We employ specific features that meet the second requirement, and other features that meet the third requirement.

The most frequently used features for gradual transition detection are color histograms. As the historic material is black and white, we use global and local luminance histograms. In Zeppelzauer et al. [2008], we use DCT coefficients and MPEG-7 edge histograms for cut detection in historic material. As the approach performs well, we also extract these features for gradual transition detection. To be invariant to object motion, we extract the luminance histograms as well as the edge histograms globally. To be more sensitive to spatial information, we extract the same features, also block-based.

4.2 Construction of the Continuity Signal

In this step we compare the feature values of successive frames to get information about signal changes, that is, the sequential continuity of the frames over time. These signal changes indicate either shot boundaries, or other significant changes in the movie, like illumination changes, and camera and object movements. We construct the continuity signal in four steps.

(1) *Normalization of the features.* We perform a min-max normalization with a minimum value of 0 and a maximum value of 1.

(2) *Self-similarity matrix calculation.* In this step we perform a pairwise comparison of the frames. The similarity matrix S is a two-dimensional matrix that contains the pairwise similarities between the feature vectors of the frames of a sequence. We employ Euclidean distance, cosine similarity, and chi-squared distance for the computation of the pairwise similarities.

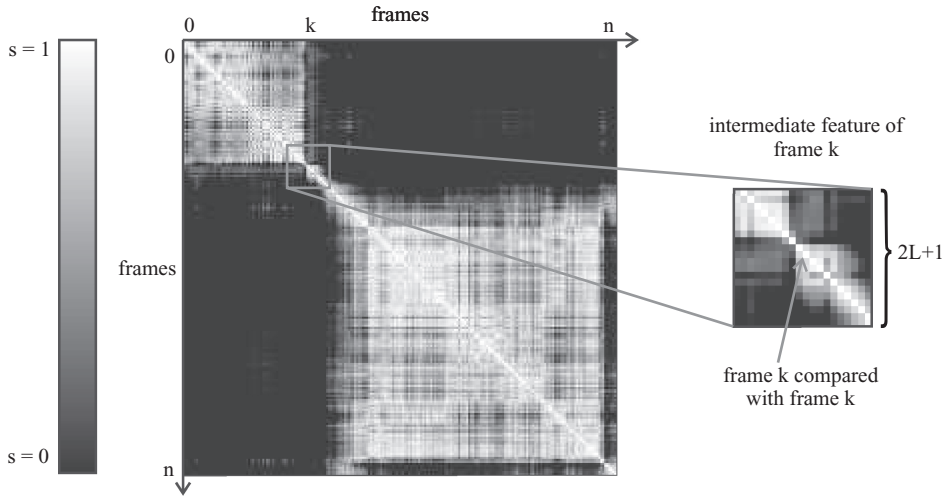

Fig. 5. Similarity matrix of a dissolve with the intermediate feature of frame k. Black areas in the similarity matrix indicate no similarity ($s = 0$) and bright areas indicate high similarity ($s \approx 0$).

(3) *Intermediate feature extraction.* The intermediate features represent the temporal neighborhood of a frame and are extracted from the similarity matrix. The intermediate feature creation kernel lag L is the number of past and future frames of a frame k that are considered for the intermediate feature. The intermediate feature of a frame k consists of the similarity matrix of the frames $k - L$ to $k + L$ (see Figure 5 for an illustration). It shows characteristic patterns for gradual transitions. As the similarity matrix is symmetric, we put one half of the intermediate feature into the intermediate feature vector. We call this a full similarity kernel. To reduce the number of feature dimensions, Cooper et al. [2007] identified the most relevant intermediate feature values to be put into the intermediate feature vector. This *greedy feature selection* results in intermediate feature vectors of lower dimensionality.

(4) *Fusion.* We combine different features to extract a maximum of information about one frame and its temporal neighborhood. For this combination, we employ two fusion strategies: early and late fusion.

Early fusion calculates a single similarity matrix from more than one feature (see Figure 6). We concatenate the different feature vectors and use the resulting vector as input for the calculation of the similarity matrix. The final intermediate feature vector for each frame is derived from this similarity matrix.

Late fusion concatenates the intermediate feature vectors instead of the features (see Figure 7). We calculate a similarity matrix for each feature. After that, we take for each frame the corresponding intermediate feature vector of each similarity matrix. Finally, we concatenate the intermediate feature vectors into a single vector.

4.3 Classification

Classification aims at identifying and labeling each frame whether or not it is part of a gradual transition. This is done by classifying the intermediate feature vector of each frame. We use a support vector machine (SVM) for training and classification [Vapnik 2000] because the dimensionality of the intermediate feature vectors tends to be high.

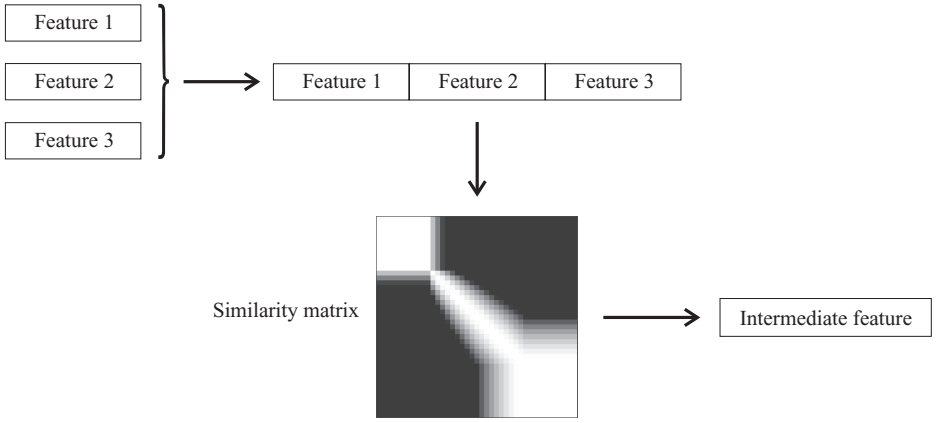

Fig. 6. Schema of feature combination with early fusion. Early fusion means that the feature vectors are first concatenated and subsequently used to compute the similarity matrix for all frames. The intermediate features are finally derived from the similarity matrix.

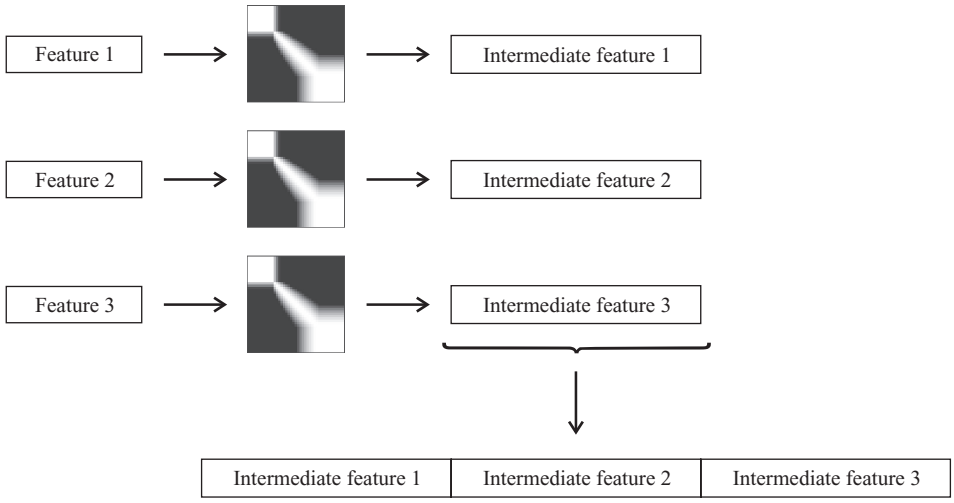

Fig. 7. Schema of feature combination with late fusion. Late fusion means that the feature vectors are used to compute the similarity matrices for all features separately. Then the intermediate features are derived from the similarity matrices. Finally, the intermediate features are concatenated.

4.4 Verification/Postprocessing

We expect many false positives and outliers in the classification results, since the historic material is of low quality. We smooth the results of the classification with a median filter in order to eliminate outliers. Furthermore, we propose Begin-End matching and KLT verification for the identification (and subsequent elimination) of false positives that occur due to camera or object motion and abrupt illumination changes.

Begin-End matching. The goal of Begin-End matching is to remove false positives. We assume that a low similarity between the frames at the beginning and at the end of a candidate gradual transition indicates a high likelihood for a correct classification. To calculate the similarity between the

beginning and the end of candidate transition, we use the similarity matrix again. We take a square of size $C \times C$ frames from the upper-right corner of the similarity matrix that represents the candidate transition. These values represent the similarity between the beginning and the end of the transition. We calculate the mean of the $C \times C$ similarity values. A low value indicates a high likelihood that a gradual transition actually occurs.

KLT verification. This postprocessing step aims at identifying false positives caused by camera and object motion. We use the KLT feature tracker to detect motion. We assume that in a sequence containing a gradual transition, all objects of the scene before the transition must disappear during the gradual transition. In case we find KLT feature points that persist through a sequence of frames classified as a gradual transition, it is likely, that the sequence is a false positive. If more than a certain number of trajectories are uninterrupted from the start frame to the end frame of a classified gradual transition, we mark the sequence as a false positive. Theoretically, we expect this threshold to work perfectly with a value of one, as one continuous trajectory already falsifies a gradual transition. In practice, higher values of the threshold give higher confidence for the falsification.

5. EXPERIMENTAL STUDY

We perform a systematic experimental study to evaluate the different parameters of the proposed method. Please refer to Sections 5.1 and 5.2 for a detailed description of the single experiments.

We employ historic material from an avant-garde film maker from the 1930s. The historic material and its artifacts are described in Section 2.2 in detail. We also evaluate the approach against the material of the TRECvid evaluation to evaluate its general validity. The TRECVid evaluation only distinguishes between cuts and gradual transitions, where a cut is a shot boundary of length zero, and a gradual transition is any other shot boundary with a length greater than zero. This is suitable for the evaluation, as we aim to detect gradual transitions independent of their type. We use the test material of the TRECVid 2006 shot boundary task. At the time of the experiments, the TRECVid 2007 evaluation test material was available as well, but the 2006 material contained more gradual transitions. The material TRECVid 2006 consists of news magazines, science news, news reports, documentaries, and educational programs.

5.1 Historic Test Data

Figure 8 shows the workflow of the experiments. We start with the evaluation of the best global and local single features and the best feature combination (using early and late fusion). For these three feature sets, we investigate the influence of different parameters on retrieval performance. In particular, we investigate the following research questions.

(1) *Which feature delivers the best results?* We evaluate each feature separately.
(2) *Which combination of features delivers the best results, and which fusion strategy performs best?* We use single features and combine them. For these combinations we evaluate early fusion, late fusion, and a mixture of both.
(3) *Which similarity measure delivers the best results?* We take the best result of a global feature and of a local feature, as well as the best result of feature combination. We evaluate the three measures Euclidean distance, cosine similarity, and chi-squared distance (sometimes referred to as *squared chi-square* distance), to find out which measure delivers the best performance with a global and a local feature, as well as with feature combination.
(4) *How does the size of L ($2L + 1$ is the kernel size for intermediate feature creation) influence the results?* We use $L = 10$ as standard. We evaluate $L = 6$ and $L = 15$ with the best global and the best local single feature and the best feature combination result.

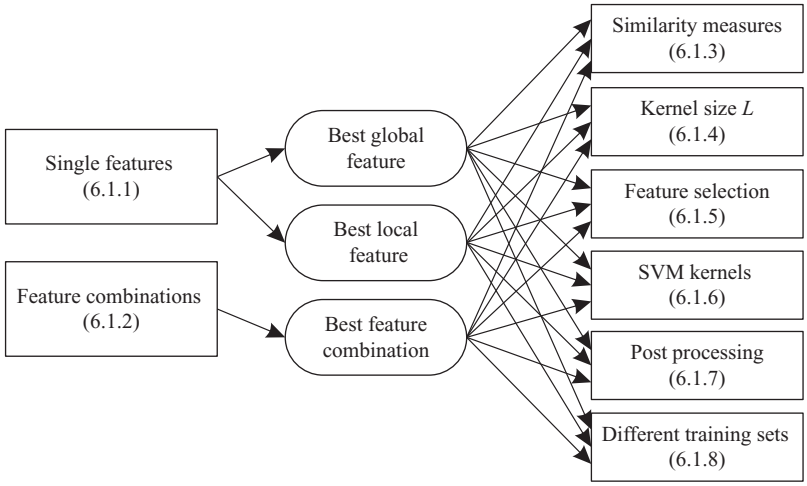

Fig. 8. Systematic overview of the experiments with historic material. We start with the evaluation of single features and feature combinations. With the best results of these experiments we evaluate further parameters (similarity measures, kernel size, feature selection, SVM kernels, postprocessing) of the method. The numbers in brackets correspond to the appropriate result sections in the article.

(5) *How do different feature selection kernels in intermediate feature creation influence the results?* We evaluate the best global and local single feature, as well as feature combinations with a greedy kernel and a full similarity kernel.

(6) *Which kernel parameters of the SVM deliver the best results?* We employ the best local feature, the best global feature, and the best feature combination to evaluate different kernel parameters for the SVM. The standard kernel used in all experiments is linear. Furthermore, we evaluate a quadratic and a polynomial kernel of third order.

(7) *What influence do the postprocessing steps have?* To answer this question, we take some of the best test runs of the previous sections and use median filtering with filter sizes from 3 to 41. Furthermore, we perform KLT verification and Begin-End matching on some results.

(8) *How do results depend on training data?* We perform all experiments with three different training datasets which contain randomly and manually selected samples, respectively.

5.2 Contemporary Test Data

We employ contemporary films as reference data to test the validity of our approach. For contemporary material, we evaluate a subset of the parameters investigated for the historic material. Based on the experience from the previous experiments, we focus on those parameters from which we expect the highest influence on retrieval performance. Figure 9 gives an overview of the experiments.

We investigate the best-performing feature and the influence of the window size of the median filter in the same way as for the historic material (see Section 5.1). We investigate the best-performing feature combination and examine the kernel lag size. Due to the suboptimal performance of early fusion in preliminary tests, we only utilize late fusion for the identification of the best-performing feature combination. For the examination of L, we employ values of $L = 6$ as the standard size and $L = 4$, as well as $L = 10$, due to the shorter lengths of gradual transitions in contemporary material (see Table I).

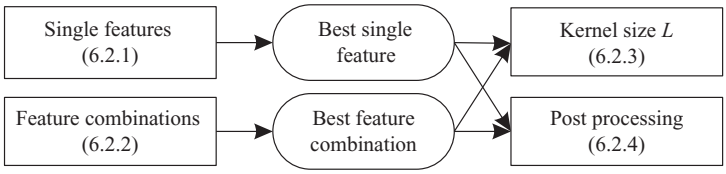

Fig. 9. To test the validity of our approach, we perform experiments on contemporary reference material from TRECVid 2006. First, we evaluate single features and feature combinations. With the best result of these experiments, we evaluate further important parameters of the proposed method. The numbers in brackets correspond to the appropriate result sections.

5.3 Evaluation

For the historic material, we use two randomly selected training data sets and one manually selected training data set to perform the SVM training in the experiments. The dependence of the results on the training data is discussed in Section 6.1.8.

For the validation of our approach with contemporary material, we use one randomly selected training data set to train the SVMs in the experiments.

We design the validation for historic and contemporary material to fit the TRECVid evaluation criteria. For this task, each participating group can submit up to 10 runs, that is, 10 detector variants to the evaluation. In our case, the different detector variants correspond to differently trained SVMs. We compare the best result (i.e., the best-performing SVM model) to the best results of the TRECVid evaluation.

We use frame recall fr and frame precision fp as performance measures in the evaluation. We combine both fr and fp to obtain the overall $f1$ measure. The usage of these measures makes our results fully comparable with the TRECVid results.

$$fr = \frac{\#(frames\ correctly\ assigned\ to\ gradual\ transitions)}{\#(frames\ belonging\ to\ gradual\ transitions\ in\ ground\ truth)} \tag{1}$$

$$fp = \frac{\#(frames\ correctly\ assigned\ to\ gradual\ transitions)}{\#(all\ retrieved\ \ frames)} \tag{2}$$

$$f1 = 2 \cdot \frac{fp \cdot fr}{fp + fr} \tag{3}$$

6. RESULTS

6.1 Historic Test Data

An overview of the systematic evaluation is given in Figure 8. First, we present and discuss the usage of different single features and the combination of features with different fusion strategies (see Sections 6.1.1 and 6.1.2). Table II contains a description and the abbreviation of the features we extract from each frame. We carry out all these feature-related experiments with the same training data set. We utilize the chi-squared distance as a measure to calculate the similarity matrices. We extract the intermediate features with a kernel lag $L = 10$. We employ full-similarity kernels for the concatenation of the intermediate feature vector. We train the SVM with a linear kernel and skip postprocessing (verification).

Second, we vary different parameters of the detector (see Sections 6.1.3 to 6.1.6 and Figure 8 for an overview). We perform these experiments with the same training data set for which we extract

Table II. Features in this Study and their Abbreviations

Feature Abbrevation	Feature Description
GLH	Global luminance histogram
LH2x2, LH3x3, LH4x4	Local luminance histograms extracted from 4, 9, and 16 blocks.
GEH	Global edge histogram
EH2x2, EH3x3, EH 4x4	Local edge histograms extracted from 4, 9, and 16 blocks.
DCT	Local DCT coefficients

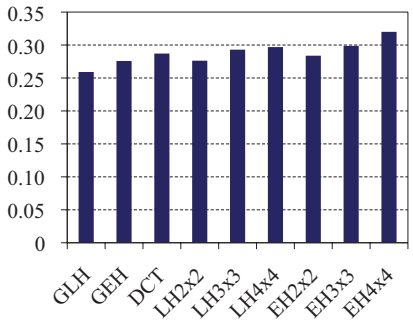

Fig. 10. Performance of single features for historic material in terms of the f1 measure. The best-performing single feature is the block-based edge histogram with 16 blocks (EH4x4).

three different feature sets from the previous experiments: we employ the best global feature, the best block-based feature, and the best feature combination. Additionally, we evaluate postprocessing (see Section 6.1.7), employing the best run of the previous experiments. Finally, we investigate the influence of the training data on the results in Section 6.1.8. We compare the experimental results of the training data set we used so far with the experimental results of two additional (training) data sets.

6.1.1 *Single Features.* The experiments with features for gradual transition detection in historic test data show that the local edge histogram with 16 blocks (EH4x4) performs best (see Figure 10). We observe, that block-based features of the same category perform better than global features of this category, and that more blocks equal better performance. This applies to the category of luminance histograms (GLH, LH2x2, LH3x3, and LH4x4) as well as the category of edge histograms (GEH, EH2x2, EH3x3, and EH4x4). In our experiments, the texture-based edge-histograms perform better than the luminance-based histograms, which contradicts earlier findings based on contemporary material [Yuan et al. 2007; Cooper et al. 2007].

6.1.2 *Feature Fusion.* Figure 11 shows the results of the feature combinations compared to the best single feature. Early fusion decreases quality in most of the cases. The combination of features with late fusion tends to improve results. The combination of all features performs best. All three late fusion combinations, which perform significantly better than the best single feature, utilize DCT in combination with at least one local feature. We conclude that the late fusion combination of DCT with a local and a global feature improves detection quality compared to the employment of single features.

6.1.3 *Similarity Measures.* Table III summarizes the results of different distance measures. We observe that the employment of the chi-squared distance is superior to the Euclidean distance and the cosine similarity with all evaluated features and feature combinations. The Euclidean distance performs comparably to the chi-squared distance with single features (GEH and EH4x4). The cosine

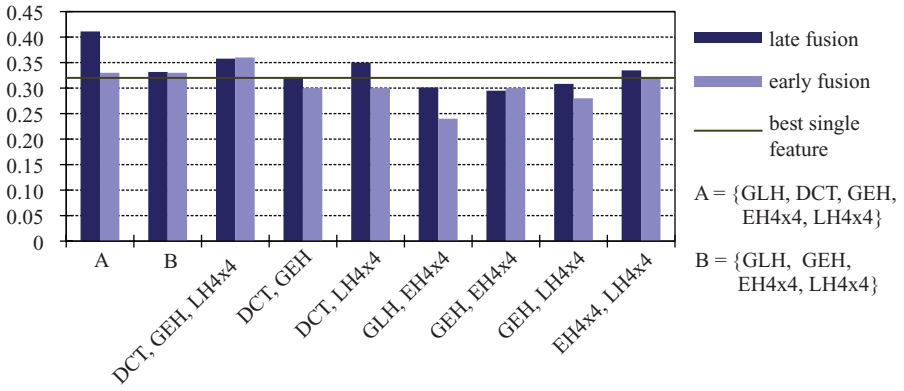

Fig. 11. Performance of features combinations with different fusion strategies in terms of the f1 measure. In the majority of cases, late fusion yields better performance than early fusion. Note that feature combination does not automatically yield better performance in relation to the best single feature.

Table III. Different Parameters for the Gradual Transition Detector

	Distance Measures			Kernel Lag L			Feature Selection		SVM Kernel		
Features	chisq	euc	cos	6	10	15	normal	greedy	linear	quad	poly
GEH	0.284	0.282	0.239	0.262	0.284	0.240	0.284	0.316	0.284	0.253	0.067
EH4x4	0.319	0.317	0.293	0.294	0.319	0.319	0.319	0.305	0.319	0.332	0.379
Late Fusion	0.411	0.385	0.375	0.304	0.411	0.478	0.411	0.343	0.411	0.337	0.239

The values are f1 values. The features we combine with late fusion are GLH, DCT, GEH, EH4x4 and LH4x4. The kernel size is calculated $2L + 1$.

similarity performs significantly worse in all three experiments. We conclude that the chi-squared distance is the best distance measure in our investigation.

6.1.4 *Intermediate Feature Kernel Size L.* We observe, that a larger kernel size for the creation of the intermediate features generally leads to a better result (see Table III). A kernel side length of 21 performs better in any case than a kernel side length of 13. This is especially true for the best feature combination with late fusion. In this case, a kernel side length of 13 results in $f1 = 0.304$, and a kernel side length of 21 results in $f1 = 0.411$. This equals an improvement of more than 25%. The increase of the kernel side length from 21 to 31 results in a significantly better performance with the late fusion run. With the employment of the best single feature (EH4x4), we observe the same retrieval performance for kernel side lengths of 21 and 31. In contrast to the local feature (EH4x4), the performance of the best global feature drops with increased kernel side length. However, in most cases a larger kernel side length results in better detector performance. We observe that a kernel side length larger than 21 is only useful if we include the information of more than one feature in the intermediate feature vector, as is the case in late fusion. We conclude, that for large L, late fusion of several features is beneficial, since it increases the information content in the intermediate feature vectors.

6.1.5 *Intermediate Feature Selection.* We observe that the experiments with greedy feature selection as proposed by Cooper et al. [2007] do not show a clear picture (see Table III). While the result with the best performing single feature (EH4x4) is relatively stable, the results for feature combinations (late fusion) are significantly worse when employing the greedily selected intermediate features. Only for the worst-performing single feature (GEH), greedy feature selection outperforms the full similarity kernel. We conclude that greedy feature selection is not appropriate for for the historic material.

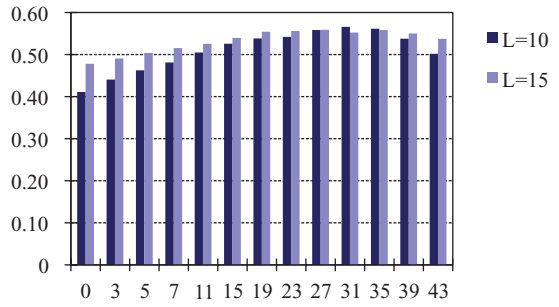

Fig. 12. F1 measure of the best result with two kernel lags $L = 10$ and $L = 15$ with a median filter applied. The application of the median filter improves the results because the filtering step reduces noise introduced by the artifacts in the historic material.

6.1.6 *SVM Kernels.* In two of three experiments the linear SVM kernel performs best (see Table III). The linear kernel is outperformed by the polynomial one only in the case of the best single feature (EH4x4) by 4%. Since a linear kernel is more stable over all experiments, we conclude, that a linear kernel suits our method best.

6.1.7 *Postprocessing.* Figure 12 shows two experiments with the same parameters, except the intermediate feature creation kernel size. The larger kernel causes better results prior to median filtering (see also Section 6.1.4). The application of a median filter with different window sizes greatly improves results in both test runs. However, the gain in performance is higher with the smaller kernel. The best median filtered results of both runs are equally good. We assume that the median filter is a possible substitute for a larger intermediate feature-creation kernel, as the median filter as well as a larger kernel compensate for the distortions (artifacts) in the historic material. In both cases, the median filter is most effective with a window size of 31 frames. We assume that this value correlates with the median length of the gradual transitions in the historic material (see Table I). The Begin-End matching and the KLT-verification do not improve retrieval performance (see Section 4.4).

6.1.8 *Dependence on Training Data.* We perform all experiments with three different training data sets. Two of them employ randomly selected training data. We evaluate whether the features deliver comparable results for both sets, and thereby whether the performance is independent from the training data set. The third training set is manually selected and is made to contain frames from *all* gradual transitions types. With the third training data set, we evaluate whether a manual optimized selection of the training data outperforms the random selection.

We observe that the results are not independent from the training data set. However, the results are consistent for the feature combinations. The three best feature combinations we identified in Section 6.1.2 are the three best combinations in each of the experiments. We observe that a manual selection of training data (which includes examples of all gradual transition types) does not improve the method's performance (see Figure 13).

6.2 Contemporary Test Data

The main goal of the experiments using contemporary material is to prove the general validity of our approach. Table IV shows the comparison of the results of the best-performing model (a linear SVM trained with the feature GLH and with $L = 6$) with those of the TRECVid 2006 shot boundary detection task. The results show that our approach performs comparably well to other approaches. This fact is noteworthy, since the method has not been optimized for contemporary material.

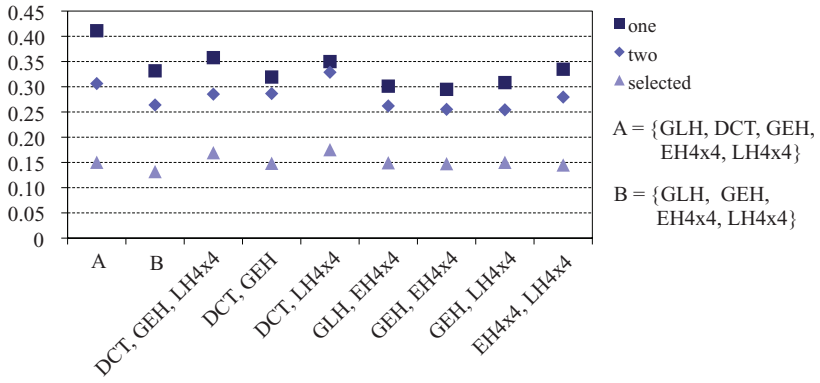

Fig. 13. This figure shows the performance (in terms of the f1 measure) for three different training data sets. The training data influences the method's performance to a certain degree. However, the best feature combinations from Section 6.1.2 consistently yield the best results.

Table IV. Contemporary Material Best Result Compared with TRECVid Results

Approach	FP	FR	F1
TRECVid 2006 best unified	0.803	0.87	0.835
TRECVid 2006 mean of all unified	0.692	0.786	0.722
Our approach	0.515	0.619	0.562
TRECVid 2006 worst nonzero unified	0.324	0.806	0.462

Additionally to testing the general validity, we aim at showing the differences in the needs of historic and contemporary material. While we use the complete TRECVid material for the validity investigation, we only use a part of the material for the following investigations to reduce computation time. We present the performance of single features, the performance of late feature fusion, the influence of the kernel size, and of postprocessing.

6.2.1 *Single Features.* Figure 14 shows the performance of single features. We observe, that the global variants of the features perform better than the block-based variants in most of the cases. The best-performing global, as well as the best-performing block-based feature, is the histogram based on the green color channel (GGH, GH4x4). In both cases, the method's performance with the luminance histograms (GLH, LH4x4) and the red color channel histograms (GRH, RH4x4) is only marginally lower than the performance with the green color channel histograms. We assume that this performance difference is explained by the proportion of red, green, and blue in the luminance Y:

$$Y = 0.299R + 0.587G + 0.114B \tag{4}$$

The green channel contains most of the luminance information; the red and green channel combined contain almost 90%. From the experiments, we observe that the color information seems to be of limited importance, since the U and V histograms alone perform poorly. In contrast to the experiments with historic material, the DCT coefficients and the edge histograms (GEH, EH4x4) perform poorly as well.

6.2.2 *Feature Fusion.* We focus on late fusion, since early fusion yielded suboptimal results in preliminary tests. Figure 15 depicts the results of late fusion with contemporary material. We observe that no feature combination improves the result compared to the best single feature (GGH). The feature combination that is closest in performance contains the best single feature in combination with the third-best single feature (GRH). The assumption regarding the limited importance of color

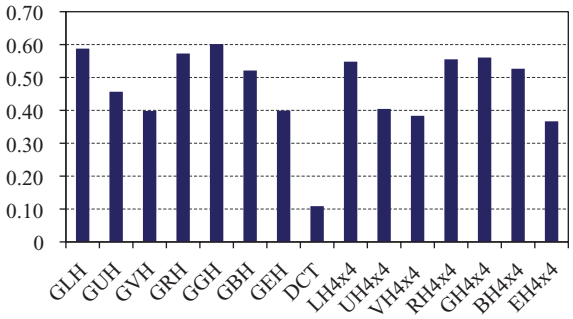

Fig. 14. Single features performance (f1 measure) with contemporary material. Additionally to the features employed for historic material (see Table II), we extract global and block-based YUV and RGB histograms. The corresponding feature abbreviations contain U,V,R,G or B instead of L (for luminance).

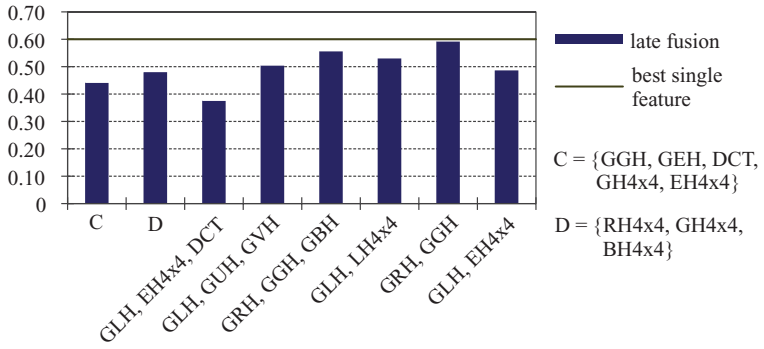

Fig. 15. For contemporary material, the best feature combination performs slightly worse than the best single feature in terms of the f1 measure; see Figure 14 for feature abbreviations.

information (see Section 6.2.1) is supported, as the combination of the global R, G, and B histograms (GRH, GGH, GBH) yields worse results than the single features GLH and GLG. We conclude that a combination of features has no benefit on the method's performance. This finding contradicts findings of earlier studies [Yuan et al. 2007; Cooper et al. 2007].

6.2.3 *Kernel Size.* The standard L has a value of 6. The experiment with the global luminance histogram resulted in $f1 = 0.588$ (see Figure 14). We conduct two further experiments with values for L of 4 and 10. The resulting $f1$ values are 0.570 and 0.565. We observe, that the kernel lag L has only a small influence on the result. Even the smallest kernel performs well. We assume that this is due to the short mean duration of the gradual transitions in contemporary material (see Table I).

6.2.4 *Postprocessing.* Figure 16 shows the application of a median filter with different window sizes on the result for the contemporary material. In contrast to the application of the median filter on historic material, we achieve no performance improvement. The contemporary material does not contain artifacts usually found in historic material. This absence of artifacts results in an undistorted signal, hence no filtering is required.

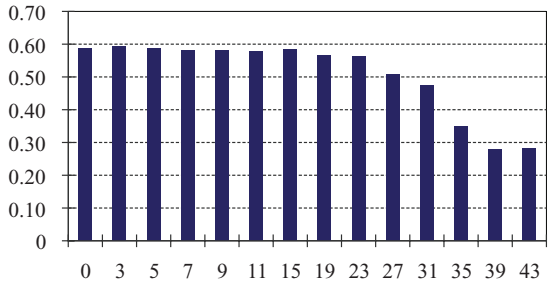

Fig. 16. The application of the median filter on results for contemporary material has very limited effect on the f1 measure. Contemporary material does not contain artifacts that introduce noise. Therefore no filtering is required.

7. CONCLUSION

We propose a method for gradual transition detection in historic material. We evaluate the proposed method with historic material of low quality and employ contemporary material of high quality as a reference. The most important lessons learned are the following.

(1) Due to flickering and brightness changes in the historic material, additional features compared to contemporary material are required. These additional features are texture-based.
(2) If we use a single feature, a local texture feature (EH4x4) performs better than any luminance feature.
(3) The combination of more than one feature with late fusion significantly improves the results. Late fusion performs better than early fusion.
(4) For historic material, a larger intermediate feature size is necessary than for contemporary material because of the longer gradual transitions in historic material. The best results for historic material are obtained with a size of 31×31 ($L = 15$), whereas for contemporary material a size of 13×13 ($L = 6$) yields the best results.
(5) The smoothing of the results with a median filter with a filter size close to the mean length of gradual transitions improves the results for historic material by 37%, whereas it hardly improves the results of contemporary material. We assume that the significant improvement is caused by the removal of outliers in the classification results. The median filter eliminates the outliers that are introduced by artifacts in the historic material.

The two other postprocessing steps we test do not improve results. The analysis of false positives shows that motion in high contrast scenes is a major problem. For example, a large black object moves in front of bright background. The postprocessing steps in literature contain temporal heuristics and SIFT features. We learn that a simple temporal median outperforms more complex approaches based on KLT trajectories and Begin-End matching.

REFERENCES

BESCOS, J., CISNEROS, G., MARTINEZ, J. M., MENENDEZ, J. M., AND CABRERA, J. 2005. A unified model for techniques on video-shot transition detection. *IEEE Trans. Multimedia 7,* 2, 293–307.

COOPER, M., LIU, T., AND RIEFFEL, E. 2007. Video segmentation via temporal pattern classification. *IEEE Trans. Multimedia 9,* 3, 610–618.

HANJALIC, A. 2002. Shot-boundary detection: unraveled and resolved? *IEEE Trans. Circuits Syst.Video Technol. 12,* 2, 90–105.

KAWAI, Y., SUMIYOSHI, H., AND YAGI, N. 2007. Shot boundary detection at TRECVID 2007. In *TREC Video Retrieval Evaluation Online Proceedings*. NIST, Gaithersburg, MD.

LIENHART, R. 2001. Reliable transition detection in videos: A survey and practitioners guide. *Int. J. Image Graph. 1*, 469–486.

LIU, Z., GIBBON, D., ZAVESKY, E., SHAHRARAY, B., AND HAFFNER, P. 2006. AT&T research at TRECVID 2006. In *TREC Video Retrieval Evaluation Online Proceedings*. NIST, Gaithersburg, MD.

LIU, Z., ZAVESKY, E., GIBBON, D., SHAHRARAY, B., AND HAFFNER, P. 2007. AT&T research at TRECVID 2007. In *TREC Video Retrieval Evaluation Online Proceedings*. NIST, Gaithersburg. MD.

SMEATON, A. F., OVER, P., AND DOHERTY, A. R. 2010. Video shot boundary detection: Seven years of TRECVid activity. *Comput. Vision Image Understand. 114,* 4, 411–418.

SMEATON, A. F., OVER, P., AND KRAAIJ, W. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR'06)*. ACM, New York, 321.

VAPNIK, V. 2000. *The Nature of Statistical Learning Theory* 2nd Ed. Springer, Berlin.

VERTOV, D. 1924. *Kinoglaz*. Collection Austrian Film Museum, Vienna, Austria.

YUAN, J., LI, J., LIN, F., AND ZHANG, B. 2005. A unified shot boundary detection framework based on graph partition model. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (Multimedia'05)*. ACM, New York, 539–542.

YUAN, J., WANG, H., XIAO, L., ZHENG, W., LI, J., LIN, F., AND ZHANG, B. 2007. A formal study of shot boundary detection. *IEEE Trans. Circuits Syst.Video Technol. 17,* 2, 168–186.

YUAN, J., ZHENG, W. J., CHEN, L., ET AL. 2004. Tsinghua University at TRECVID 2004: Shot boundary detection and high-level feature extraction. In *NIST Workshop of TRECVID*. NIST, Gaithersburg, MD.

ZEPPELZAUER, M., MITROVIĆ, D., AND BREITENEDER, C. 2008. Analysis of historical artistic documentaries. In *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE Computer Society, Los Alamitos, CA, 201–206.

ZHANG, H., KANKANHALLI, A., AND SMOLIAR, S. W. 1993. Automatic partitioning of full-motion video. *Multimedia Syst. 1,* 1, 10–28.

Received March 2011; revised April 2011