A Study of Gradual Transition Detection in Historic Film Material

Markus Seidl
St. Pölten University of Applied Sciences
Matthias-Corvinus Straße 15
A-3100 St. Pölten
markus.seidl@fhstp.ac.at

Matthias Zeppelzauer
Vienna University of Technology
Favoritenstraße 9-11/188/2
A-1040 Vienna
mzz@ims.tuwien.ac.at

Christian Breiteneder
Vienna University of Technology
Favoritenstraße 9-11/188/2
A-1040 Vienna
cb@ims.tuwien.ac.at

ABSTRACT
The detection of gradual transitions focuses on two types of approaches: unified approaches, i.e. 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 our 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 colour and luminance features.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]:–*Indexing Methods*;

H.5.1 [**Multimedia Information Systems**]:–*Video*;

General Terms
Algorithms, Experimentation.

Keywords
Shot Boundary Detection, Gradual Transition Detection, Cultural Heritage.

1. INTRODUCTION
Today, film archives provide large amounts of (digitized) films. The access to these films is difficult because they are neither segmented nor annotated.

Automatic shot boundary detection is a first step to enable nonlinear access and browsing of these videos.

eHeritage'10, October 25, 2010, Firenze, Italy.

A shot is defined as a continuous sequence of frames which have been captured in one camera action. The transition between shots can be abrupt or gradual, and is therefore defined as cut or gradual transition (GT).

The National Institute of Standards and Technology (NIST) started the TRECVid benchmark for content-based retrieval in 2001 [1]. Since then, a large amount of test footage for shot boundary detection became available and enabled the objectice comparison of different approaches. Since 2005, the detection of cuts can be considered as solved [2], since 2008 the detection of GTs is also declared as solved [1][3].

For historic material, we agree with that in the case of CUTs. We achieved satisfactory results in previous work on CUT detection in historic material in [4]. However, we disagree for the case of GT detection in historic material. If shot boundary detection (SBD) is applied to historic material, we have novel challenges to solve. This type of material has special properties that we have to take into account. For example artifacts like scratches, fungus, flicker and shaking as well as complex and long gradual transitions. One might expect that a priori film restoration solves most of these problems. However, film restoration is a time consuming and expensive process that 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.

To our knowledge no research about the detection of gradual transitions in historic material has been published so far. We perform a broad experimental study to gain insight in how the characteristics of historic material influence the steps of the GT detection process.

This paper is organized as follows: In Section 2, we define the problem and present related work. In Section 3, we explain our method for GT detection in historic material. In Section 4 we describe the experimental study. In Section 5 we present the most important results, and in Section 6 we draw conclusions.

2. PROBLEM DESCRIPTION AND MATERIAL
GT detection for contemporary video material is a well investigated problem. We briefly state the problems of GT detection in general and explain the specific problems in the context of historic material. Then we summarize promising approaches known from literature.

2.1 Problem Description

The main problems with GT detection are: (i) many different GT types exist, (ii) the GTs have varying lengths, and (iii) object- and camera movement produce signal changes comparable to those of a GT [2][5][6]. These three main problems are described well by Yuan et al. [7]. Additionally to these three problems, two specific problems occur with historic material: first, the material includes different and longer GTs (see Table 1) and second, the material contains many artifacts that interfere with GT detection.

Table 1: Gradual transition lengths in historic and contemporary material.

Material	#GT types	#Frames / GT		
		Min	Max	Mean
Historic	8	6	134	30.7
TRECVid	3	1	107	9.3

2.2 Material

The historic material we employ is Kinoglaz by Dziga Vertov. It was produced in 1924 and has a runtime of 78 minutes at 18fps. It is black and white consequently. Features that rely on colour information are not usable. The employed copy is several decades old. This causes two sources of defects: long storage of the movie, e.g. fungus and dirt; and technical limitations in the 1920ies, e.g. flickering.

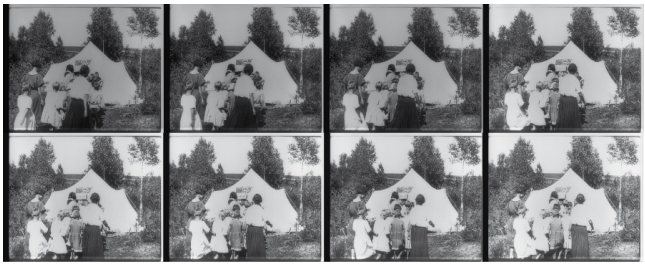

Figure 1: Flicker in Kinoglaz frames 20808 to 20815.

The types of defects are: *(i) Flicker:* Due to manual film transport in old cameras the brightness of the films is unsteady (see Figure 1). *(ii) Image vibrations:* Due to the shrinking of the film, the images vibrate (see Figure 2). *(iii) Degraded contrast:* Due to the fatigue of the material the contrast of films is degraded.

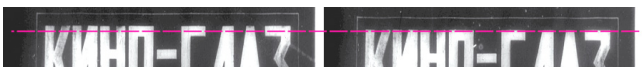

Figure 2: Image vibrations in Kinoglaz frames 6 to 9.

(iv) Scratches: The films have long vertical scratches due to mechanical problems in the old projectors. *(v) Fungus / dirt:* The films contain dirt and a visible fungus is growing on the material (see Figure 3). *(vi) Wrong exposure or development:* Unreliable light meters and unreliable camera shutters or not-standardized development under non-stable conditions caused wrong exposure or development.

Figure 3: Dirt on Kinglaz frame 60.

3. RELATED WORK

We survey and compare existing approaches in two categories: specialised approaches, i.e. one detector for each GT type, and unified approaches, i.e. one detector for all GT types. We focus on specialised approaches for dissolve detection since dissolves are the most common GTs.

The twin comparison method for the detection of dissolves [8] was one of the first GT detection methods. It utilizes accumulated inter-frame differences to achieve inter-temporal comparison longer than one frame. It uses colour histograms as features. The detection is done by thresholding.

Yuan et al. improve the twin comparison method for gradual transitions lasting longer than 5 frames [9]. The approach includes motion vectors that control a self-adapting threshold. The features used are colour histograms, difference images and motion vectors.

Kawai et al. propose a method that is based on two assumptions [10]. 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 post processing steps are heuristical verification (duration of a dissolve) and a comparison of the similarity of the begin and end frame of the dissolve candidate.

The three approaches mentioned so far are based on colour information and employ thresholds for detection. The usage of thresholds causes a lack of robustness. Yuan et al. state, that a threshold's value highly depends 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 [2].

Liu et al. delivered the best SBD performance in the 2006 and 2007 TRECVids. They propose a dissolve detector based on the change of the colour histogram variance during a dissolve [11][12]. They assume that a dissolve is a linear mixture of two shots and that therefore the change of the colour variance during a dissolve follows typical curves. These curves are detected by finite state machines. The detector extracts colour 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).

Bescós et al. introduce in [13] a unified approach for gradual transition detection which is based on the patterns that result from inter-frame comparison with different temporal distances. They use RGB colour values as features and classify by thresholding.

Yuan et al. utilize in [2] a similarity matrix with inter-frame 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 similarity matrix is calculated in lower resolution, e.g. 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 colour histograms. The classification is done with an SVM.

Cooper et al. propose in [14] a well performing SBD approach that calculates a similarity matrix containing the inter-frame similarities of all frames of a sequence. The detector uses the window in the similarity matrix around the frame as 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 colour histograms. The verification step employs temporal heuristics.

Specialised approaches are more complex, as they use a single detector for each GT type and have to merge results in an additional step. Each GT type requires a special detector, and each detector needs training. The unified approaches consist of less processing steps than the specialized ones. They are more general and therefore more transparent. As the unified approaches rely on inter-temporal comparison of more than two frames, we expect them to be more robust against distorted material. The specialised approaches have a better detection performance. As unified and specialised approaches rely mainly on colour values and histograms as features, we expect both to be equally b/w incompatible.

The approach of Bescós et al. relies on the pattern of inter-temporal comparison of all frames with one frame. We expect a lack of robustness for distorted material. The approaches of Yuan et al. and Cooper et al. rely on similarity matrices that compare all frames of a sequence with each other. Therefore we expect these two approaches to be superior to the approach of Bescós et al. Both approaches are tested against the TRECVid SBD task material and perform comparably well.

We adapted the approach of Cooper [14] to the detection of abrupt transitions in historic material in [4]. The approach yields satisfactory results for CUT detection. In this paper, we extend our work to the detection of gradual transitions in historic material.

4. PROPOSED METHOD

According to Yuan et al. [7] a SBD system consists of three processing steps: (i) visual content representation, (ii) construction of the continuity signal and (iii) classification. We add a fourth step of (iv) verification, as we perform a post processing 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 requirements special to our approach are: (i) invariances caused by flickering frames, scratches, fungus 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. We evaluate in the study, whether features meeting requirement (ii) or features meeting requirement (iii) or a combination of both suits GT detection best.

The most frequently used features for GT detection are colour histograms. As the historic material is black and white, we use global and local luminance histograms. In [4], 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 GT 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 block-based. Table 2 contains a description of the features we extract from each frame.

Table 2: Features extracted from each frame.

Feature abbreviation	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

4.2 Construction of the Continuity Signal

In this step, we compare the features of successive frames to get information about signal changes, i.e. the sequential continuity of the frames over time. These signal changes indicate either shot changes, or other significant changes in the movie, like illumination changes or camera and object movement.

We construct the continuity signal in four steps.

(i) Normalisation of the features.

(ii) Similarity matrix calculation: In this step we perform a pair-wise comparison of the frames. The similarity matrix S is a two dimensional matrix that contains the pair-wise 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 pair-wise similarities.

(iii) Intermediate feature extraction: The similarity matrix is the basis for the intermediate features that represent the temporal neighborhood of a frame. The intermediate feature of a frame k consists of the similarity matrix of the frames k-L to k+L (see Figure 4). It shows characteristic patterns for gradual transitions. The extraction of similarity matrices and intermediate features is described in more detail by Cooper et al. [14].

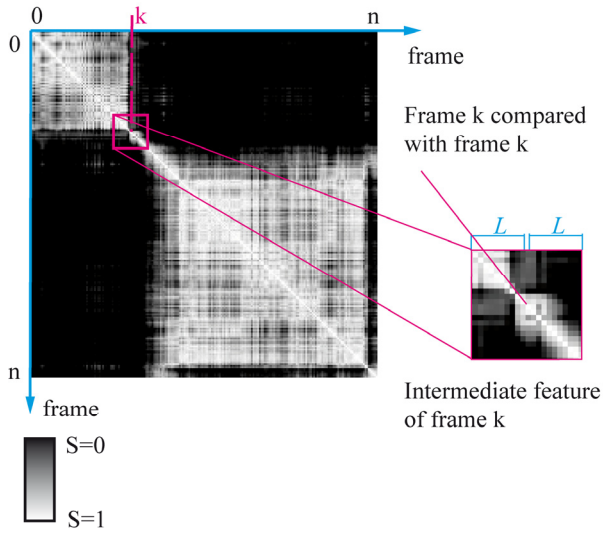

Figure 4: Similarity matrix of a dissolve with the intermediate feature of frame k.

(iv) 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. 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. 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 labelling each frame whether or not it is part of a gradual transition. This is done by classifying the intermediate features of each frame. The dimensionality of the intermediate feature vectors tends to be high. We use a Support Vector Machine (SVM) for training and classification [15].

4.4 Verification / Post processing

Since our material is of low quality, we expect many false positives and outliers in the classification results. To eliminate outliers we first smooth the results of the classification with a median filter.

Then, in order to identify false positives that occur due to camera or object motion and abrupt illumination changes, we propose two verification steps for the removal of false positives: Begin-end matching and KLT verification.

Begin-End matching: The goal of begin-end matching is to remove false positives. We assume that a high similarity between the frames at the beginning and at the end of a sequence indicates a small likelihood for a shot boundary in the sequence. To calculate the similarity between the beginning and the end of the sequence, we use the similarity matrix again. We take a square of size C*C in the upper right or in the lower left corner of the similarity matrix and calculate the mean of the similarity values. A high value means a high likelihood that no shot transition occurs in this sequence.

KLT verification: This post processing 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 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 already one continuous trajectory falsifies a gradual transition. In practice, higher values of the threshold give more confidence of the falsification.

5. EXPERIMENTAL STUDY

We perform a systematic experimental study with numerous tests. We study: (i) different features, combination of features and fusion strategies; (ii) the dependency on the training data; (iii) the performance of the post processing steps; and (iv) the performance of our approach for contemporary reference data.

We employ historic material from an avant-garde film maker. The historic material and its artefacts are described in Section 2. We test our approach also against material of the TRECvid evaluation to evaluate the general validity of our approach. 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 zero. This is suitable for our approach, as we aim at finding gradual transitions independent of their type.

We use the test material of the TRECVid 2006 shot boundary task. At the time of our experiments, the TRECVid 2007 evaluation test material was available as well, but the 2006 material contains more gradual transitions. The material consists of news magazines, science news, news reports, documentaries and educational programs.

6. RESULTS

We use frame recall *fr* and frame precision *fp* as measures for our results. We combine both to the widely used *f1* measure. The usage of these measures makes our results fully comparable with the TRECVid results.

6.1 Single Features and Feature Fusion

The following results are all from one training data set. The SVM is trained with a linear kernel.

As it performs best in all our experiments, we use the Chi-squared distance as measure to calculate the similarity matrices. We present the results prior to any verification. The experiments with features for gradual transition detection in our historic test data shows that the local edge histogram with 16 blocks performs best (see Figure 5).

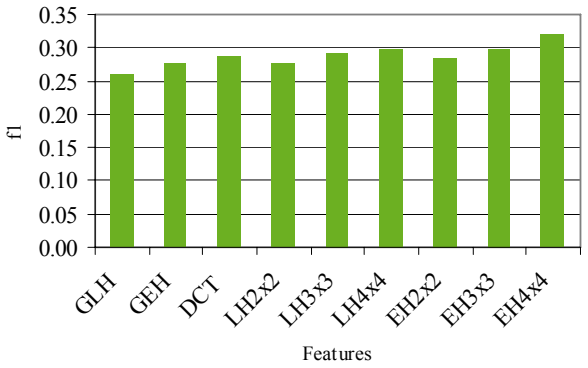

Figure 5: Performance of single features.

Figure 6 shows the results of the feature combinations compared with the best single feature. Early fusion decreases result quality in most of the cases. The combination of features with late fusion sometimes improves results. The combination of all features performs best. All three late fusion feature combinations, that perform significantly better than the best single feature, utilize DCT in combination with other features that include 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.

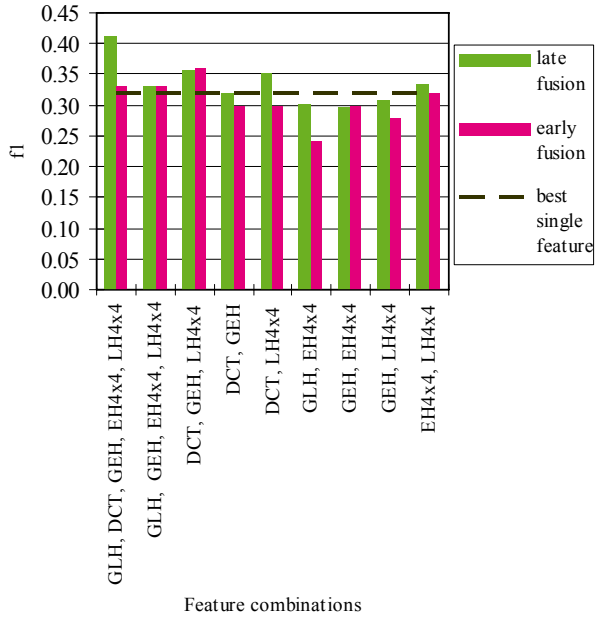

Figure 6: Performance of feature combinations.

6.2 Dependence from Training Data

We perform all our experiments with three different training data sets. Two of them employ randomly selected training data. We evaluate first, whether the features deliver comparable results in both sets, and second, whether the results are dependent from the training data set. With the third training data set we evaluate, whether the usage of gradual transition frames of all gradual transition types as training data improves result. We select the gradual transitions frames of this training data set manually.

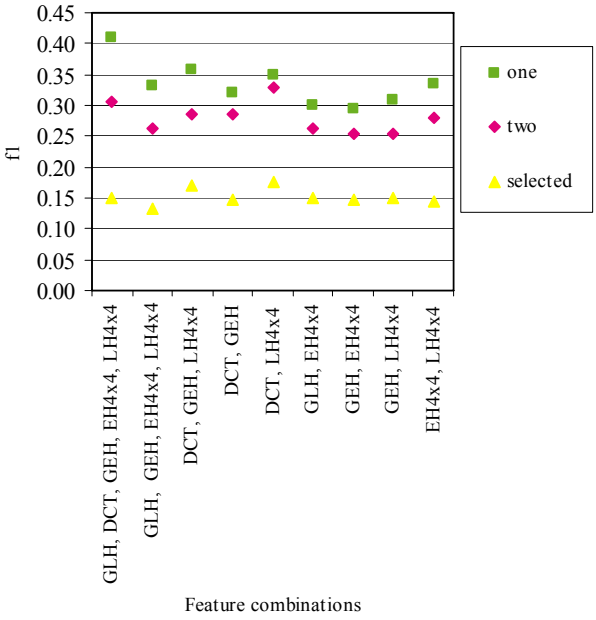

Figure 7: Dependence of training data.

We observe that the results are not independent from the training data set. However, the results are usable for the comparison of the performance of the feature combinations. The three best feature combinations we identified in Section 5.2 are the three best combinations in each of the experiments. We observe that a manual selection of training data that consciously contains all gradual transitions types is no improvement for the results (see Figure 7).

6.3 Post Processing

The post processing with a median filter greatly improves results (see Figure 8). The median filter is most effective with a window size of 31 frames. The begin-end matching and the KLT-verification are no improvement for the results (see Section 6).

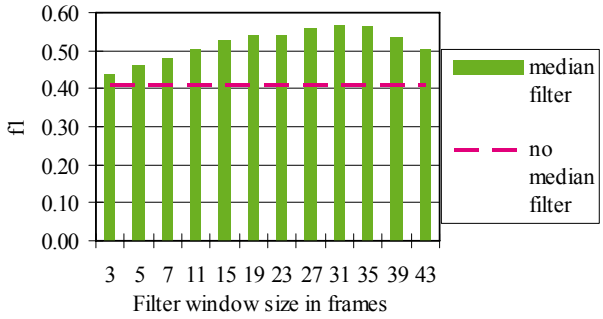

Figure 8: Median filter applied on the best result.

6.4 Contemporary Material

The aim of our test runs with contemporary material is to prove the general validity of our approach. Due to limited space we only present the best result, which is obtained by a global luminance histogram with 16 bins. In contrast to other studies, neither the combination of features nor the usage of colour features improve detection performance [7][14]. Table 2 shows the comparison of our results with those of the TRECVid 2006 SBD task. The SVM was trained with a linear kernel.

Table 3: 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 non-zero unified	0.324	0.806	0.462

The results show, that our approach performs comparably well with other approaches, especially as the method has not been optimized for contemporary material.

7. CONCLUSION

We proposed a method for gradual transition detection in historic material. We carried out experiments on historic material with low quality and contemporary high quality reference material. The most important lessons learned are:

(i) Due to flickering and brightness changes in the historic material additional features compared to contemporary material are required. These additional features are texture-based.

(ii) If we use a single feature, a local texture feature (EH4x4) performs better than any luminance feature.

(iii) The combination of more than one feature with late fusion significantly improves the results. Late fusion performs better than early fusion.

(iv) Due to the length of the gradual transitions in historic material a larger intermediate feature size than with contemporary material is necessary (see Figure 4). The best results are obtained with a size of 31x31 (L=15), in contrast to a size of 13x13 (L=6) for contemporary material.

(v) The smoothing of the results with a median filter with a filter size close to the mean length of gradual transitions improves the historic material results by 37%, whereas it hardly improves the results of contemporary material. We assume that the significant improvement is caused by the noise in the historic material that causes outliers in the classification results, which are removed by the median filter.

The two other post processing steps we test do not improve results. The analysis of false positives shows that motion in high contrast scenes, generally "black object moves in front of bright background", is the major problem. The post processing steps in literature contain temporal heuristics and SIFT features. The employment of the median filter as simple temporal heuristic yields satisfactory results, while the KLT features we used proofed suboptimal. We assume that this might be due to flicker, scratches, etc. and that the SIFT features might pose this challenge as well.

8. REFERENCES

[1] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," *Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06*, Santa Barbara, California, USA: 2006, p. 321.

[2] J. Yuan, J. Li, F. Lin, and B. Zhang, "A unified shot boundary detection framework based on graph partition model," *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 539–542.

[3] A.F. Smeaton, P. Over, and A.R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Computer Vision and Image Understanding*, vol. 114, 2010, pp. 411-418.

[4] M. Zeppelzauer, D. Mitrovic, and C. Breiteneder, "Analysis of Historical Artistic Documentaries," *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, 2008, p. 201–206.

[5] R. Lienhart, "Reliable Transition Detection in Videos: A Survey and Practitioner´s Guide," *International Journal of Image and Graphics*, vol. 1, 2001, p. 469-486.

[6] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, Feb. 2002, p. 90-105.

[7] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, "A Formal Study of Shot Boundary Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, 2007, p. 168-186.

[8] H. Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, 1993, S. 10-28.

[9] J. Yuan, W.J. Zheng, L. Chen, and others, "Tsinghua university at trecvid 2004: Shot boundary detection and high-level feature extraction," *NIST workshop of TRECVID*, 2004.

[10] Y. Kawai, H. Sumiyoshi, and N. Yagi, "Shot boundary detection at TRECVID 2007," *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[11] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray, and P. Haffner, "AT&T Research at TRECVID 2006," *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[12] Z. Liu, E. Zavesky, D. Gibbon, B. Shahraray, and P. Haffner, "AT&T Research at TRECVID 2007," *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[13] J. Bescos, G. Cisneros, J.M. Martinez, J.M. Menendez, and J. Cabrera, "A unified model for techniques on video-shot transition detection," *IEEE Transactions on Multimedia*, vol. 7, 2005, pp. 293–307.

[14] M. Cooper, T. Liu, and E. Rieffel, "Video segmentation via temporal pattern classification," *IEEE Transactions on Multimedia*, vol. 9, 2007, pp. 610–618.

[15] V. Vapnik, *The nature of statistical learning theory*, New York: Springer, 1995.