A Formal Study of Shot Boundary Detection

Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang

***Abstract*—This paper conducts a formal study of the shot boundary detection problem. First, a general formal framework of shot boundary detection techniques is proposed. Three critical techniques, i.e., the representation of visual content, the construction of continuity signal and the classification of continuity values, are identified and formulated in the perspective of pattern recognition. Meanwhile, the major challenges to the framework are identified. Second, a comprehensive review of the existing approaches is conducted. The representative approaches are categorized and compared according to their roles in the formal framework. Based on the comparison of the existing approaches, optimal criteria for each module of the framework are discussed, which will provide practical guide for developing novel methods. Third, with all the above issues considered, we present a unified shot boundary detection system based on graph partition model. Extensive experiments are carried out on the platform of TRECVID. The experiments not only verify the optimal criteria discussed above, but also show that the proposed approach is among the best in the evaluation of TRECVID 2005. Finally, we conclude the paper and present some further discussions on what shot boundary detection can learn from other related fields.**

***Index Terms*—Formal framework, graph partition model, multiresolution analysis, shot boundary detection, support vector machine (SVM).**

I. Introduction

RECENT advances in multimedia compression technology, coupled with the significant increase in computer performance and the growth of the Internet, have led to the widespread use and availability of digital videos. The rapidly expanding applications of videos have spurred the growing demand of new technologies and tools for efficient indexing, browsing and retrieval of video data. The area of content based video retrieval, aiming to automate the indexing, retrieval and management of video, has attracted extensive research during the last decade [1], [2].

Structural analysis of video is a prerequisite step to automatic video content analysis. Among the various structural levels (i.e., frame, shot, scene, etc.), shot level organization has been considered appropriate for browsing and content based retrieval [3], [4]. A shot consists of continuous frame sequences captured by a single camera action. According to whether the transition between shots is abrupt or gradual, the shot boundaries can be categorized into two types: cut (CUT) and gradual transition (GT). The GT can be further classified into dissolve, wipe, fade out/in (FOI), etc., according to the characteristics of the different editing effects [5]. Shot boundary detection (SBD), also known as temporal video segmentation, is the process of identifying the transitions between the adjacent shots. A large number of SBD methods have been proposed. In the early years, the methods are usually evaluated on a relatively small data set due to the lack of large annotated video collections. Since the year of 2001, the National Institute of Standards and Technology (NIST) has started a benchmark of content based video retrieval, i.e., TRECVID [6], in which SBD is one of the evaluation tasks. NIST provides much larger evaluation data than ever. Dozens of participants present their SBD approaches for evaluation. The practice of TRECVID has significantly promoted the progress of SBD techniques. It reveals that the identification of CUTs has been somewhat successfully tackled, while the detection of GTs still remains a difficult problem [6].

Despite the extensive research on concrete SBD techniques, little attention has been paid to the formal study of the problem. To our best knowledge, Vasconcelos *et al.* made the initial efforts to formulate the problem [7]. They developed a Bayesian formulation for the problem and extended the standard thresholding model to an adaptive and intuitive way. In [8], Lienhart identified several core techniques underlying the various SBD schemes and reviewed their roles in detecting CUTs, fades and dissolves. In [9], Hanjalic analyzed the SBD problem and identified the major issues that needed to be considered for a successful approach. Recent formal study on SBD includes [10] and [11]. Albanese *et al.* presented mathematical characterizations for most common transition effects [10]. Bescós *et al.* proposed a unified model centering on the mapping from the feature space to the space of inter-frame distances and the mapping from the distances space to the decision space [11]. This model is capable of covering most of the existing SBD techniques. These formalizations make the essence of SBD explicit, meanwhile, they identify the crucial functional components and clarify the pros and cons of the existing approaches. The formal study will inevitably guide the development of novel SBD techniques.

However, there is still some work remaining unsolved. First, we have to say that even the latest formal studies have not been advanced enough to cover the recent development of SBD techniques, especially for the methods appeared after the year of 2000. For example, all of the previous work has formalized the decision procedure of SBD as a thresholding model. Nevertheless, some recent work (e.g., [12]–[14]) can not be exactly described by metrics or thresholds, since their final decisions are not obtained by thresholding schemes but by machine learning

Manuscript received May 20, 2006; revised August 29, 2006. This work was supported by National Natural Science Foundation of China under Grants 60135010 and 60321002 and in part by the Chinese National Key Foundation Research and Development Plan under Grant 2004CB318108. This paper was recommended by Associate Editor P. Lobato Correia.

J. Yuan, H. Wang, W. Zheng, J. Li, F. Lin, and B. Zhang are with the Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China (e-mail: yuan-jh03@mails.tsinghua.edu.cn; hy-wang02@mails.tsinghua.edu.cn; idiot00@mails.tsinghua.edu.cn; lijianmin@mail.tsinghua.edu.cn; linfz@mail.tsinghua.edu.cn; dcszb@mail.tsinghua.edu.cn).

L. Xiao is with the Computer Science Department, Columbia University, New York, NY 10027 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2006.888023

methods. Some calculation methods of content discontinuity are also not covered by the previous formal study. In [12] and [13], they did not calculate the discontinuity values by by pair-wise frame comparison but by incorporating context information of the neighborhood. Second, the previous work, though identified various specific core techniques, did not evaluate these ones. The reliable evaluation and the optimal criteria for designing a specific functional component are extremely important for the practical purpose. It is time to conduct a novel formal study on SBD. Numerous techniques have been developed and several comprehensive surveys have been presented to summarize them [8]–[11], [15]. These efforts have established the foundation of the further formal study. Meanwhile, the platform of TRECVID has provided the facilities for the reliable evaluation of various techniques.

In this paper, we conduct a formal study of the SBD problem. First, we present a general formal framework for SBD techniques in the perspective of pattern recognition. Three critical techniques, i.e., the representation of visual content, the construction of continuity signal and the classification of continuity values, are identified and formulated. Second, we present a comprehensive review of the existing approaches. In the review, the representative approaches are categorized and compared according to their roles in the formal framework. Based on the comparison of the existing approaches, optimal criteria for each module of the framework are discussed. Third, with all the above issues considered, we present a unified SBD system based on graph partition model. Extensive experiments are carried out on the platform of TRECVID. The experiments not only verify the optimal criteria discussed above, but also show that the proposed approach is among the best in the evaluation of TRECVID 2005. Finally, we present some further discussions on what SBD can learn from other related fields.

The remainder of this paper is organized as follows. Section II presents a formal framework for the SBD techniques. Section III provides the review of the existing methods. The Section II and Section III focus on formally analyzing the SBD problem and identifying the major challenges while designing SBD system. As an example of the formal framework, we introduce an SBD system based on graph partition model in Section IV. In Section V, we carry out some comparative experiments on the platform of TRECVID to examine the effectiveness of the proposed system. We conclude this paper and outline the future possible directions in Section VI.

II. Formal Framework of SBD

In this section, we attempt to establish a general formal framework for SBD techniques and point out the major challenges to the framework. Video is composed of multistreams of information, i.e., audio, visual, text, etc.. However, all of the existing SBD systems recognize shot boundaries according to the transitions of visual content, except [16] which incorporated scripts of automatic speech recognition (ASR). This is mainly due to the following two reasons. First, visual content is the major information source of videos and it will yield better detection results for such structure analysis of physical level [1]. Second, the fusion of multimodalities still remains a challenge in the field of multimedia content analysis [2]. People have not found effective ways to perform combined and cooperative analysis of multimodalities in the cases of heterogeneous and even conflicting information. In this paper, we will focus on the visual aspect of videos. Nevertheless, in the framework, the visual information is abstracted by its features. It is possible to replace visual feature by other features such as audio. The evidences from multisources can be combined in the way of information fusion (e.g., the multiresolution analysis in Section IV).

A. Formal Definition of SBD

In the perspective of visual aspect, video is a kind of three-dimensional signal, in which two of them reveal the visual content in the horizontal and vertical frame direction, and the third one reveals the variations of the visual content over the time axes. Neglecting the signal variation along the horizontal and vertical dimensions, let I_t denote the tth frame, where $I_t \in Q$ and Q indicates the image space. SBD aims to temporally segment the video into some consecutive shots, i.e., uninterrupted image sequences captured by a single camera action. The basic idea of SBD approaches is to identify the discontinuities of visual content. No matter what kind of detection techniques, it consists of three core elements, i.e., the representation of visual content, the evaluation of visual content continuity and the classification of continuity values. In the following, we will introduce the formalizations of the above three modules, respectively. Note that the style of the formalizations is similar to that of Bescós *et al.* [11], but the content is distinct.

1) Representation of Visual Content: The straightforward approach to representing the visual content of each frame I_t is to utilize the image itself. A more popular alternative is to extract some kind of visual features from each frame and obtain a compact content representation. Let $V_t \in F$ denote the feature of I_t, where F is the feature space. The problem of content representation is to seek an appropriate feature extraction method, formally, to find a mapping from the image space Q to the feature space F

$$\begin{aligned} \Phi : Q &\longrightarrow F \\ I_t &\longrightarrow V_t. \end{aligned} \tag{1}$$

There are two major requirements for feature V_t as an appropriate content representation, i.e., *invariance* and *sensitivity*. Here, the invariance means that the feature is stable to some forms of content variation except shot transitions, e.g., rotation or translation of the picture. Inversely, the sensitivity reflects the feature's capability of capturing the details of visual content. Generally speaking, the rougher the feature is, the stronger the invariance is. For example, color coherent vector is more sensitive than color histogram, since color coherent vector is a refinement of color histogram which incorporates the spatial information of color distribution [17]. The sensitivity is a reverse aspect of invariance. The more details the feature can capture, the more sensitive it is, since the feature can even reflect the tiny changes of visual content. With the invariance, the feature within shots remains relatively stable, while with the sensitivity, the feature between different shots exhibits considerable change. The tradeoff between invariance and sensitivity must be taken into account to achieve a satisfying detection performance.

2) Construction of Continuity Signal: The common practice of identifying the transitions between shots is to first calculate the continuity (similarity) or discontinuity (distance) values of adjacent features (In this paper, we adopt the continuity signal, which is just an inverse of the discontinuity signal of the related literature [8], [9]). In this way, the visual content flow is transformed into a 1-D temporal signal. In the ideal situation, the continuity signal within the same shot always keeps large magnitudes, while drops to low values surrounding the positions of shot transitions. Unfortunately, the temporal signal obtained by inter-frame comparison of features is not always stable enough to various disturbances such as abrupt illumination variation and large object/camera movement. A better way is to not only consider inter-frame variations but also incorporate the variations within the neighborhood of the particular position, i.e., contextual information. Formally, let S denote the space of continuity values and s_t be the content continuity between V_t and V_{t+1}. The procedure of continuity signal calculation is to construct a mapping from the Cartesian product of feature space to the continuity value space

$$\begin{aligned}\Theta : F^{2\times d} &\longrightarrow S\\ A_t^d &\longrightarrow s_t\end{aligned} \tag{2}$$

where $A_t^d = (V_{t-d+1}, \ldots, V_t, V_{t+1}, \ldots, V_{t+d})$ and d denotes the radius of the involved neighborhood when calculating the content continuity between V_t and V_{t+1}. In the early approaches, d usually equals 1, that is, only the inter-frame continuity is evaluated. While in some recent work, values of $d > 1$ are adopted to incorporate contextual information as reviewed in Section III. Note that in (2) we have assumed the video content flow is transformed to a 1-D continuity signal. To our best knowledge, all of the existing approaches conform to the assumption. However, this is not the only solution. We can design some vectorial continuity signals to reflect content variation. In Section IV, we will introduce a module named *Construction of Multiresolution Graph* as an example. The following presentation will continue using the assumption of 1-D signal for the convenience of explanation.

3) Classification of Continuity Values: Given the 1-D temporal signal of content variation, the final critical issue is to classify the boundaries from the nonboundaries or identify the types of the transitions. This procedure is also a mapping but from the Cartesian product of continuity value space S to the decision space W. Let $w \in W$ denote the type of transitions (or nonboundary) between I_t and I_{t+1}, the mapping of decision can be indicated as

$$\begin{aligned}\Psi : S^{2\times r+1} &\longrightarrow W\\ B_t^r &\longrightarrow w\end{aligned} \tag{3}$$

where $B_t^r = (s_{t-r}, \ldots, s_t, s_{t+1}, \ldots, s_{t+r})$ and r is the radius of the neighboring continuity values required by the classifier. In most cases, $r = 0$ is adopted, that is, only the amplitude of s_t is used to determine whether a transition occurs between I_t and I_{t+1}. In some recent approaches [12], [13], the nearby temporal pattern of the continuity signal is considered while judging the presence or absence of shot transition, i.e., $r > 0$. In the perspective of Bayesian decision theory, the optimal mapping can be obtained as

$$\Psi\left(B_t^r\right) = \arg\max_{w\in W} P\left(w|B_t^r\right) \tag{4}$$

in which $P(w|B_t^r)$ indicates the posterior probability of the transition type being w given the observation B_t^r. The mapping obtained according to (4) is the so-called minimum error rate classifier. How to construct a mapping with the minimum error rate is the core problem of machine learning theory. Generally speaking, there are two different ways to model Ψ, namely generative and discriminative classifiers. In generative classifiers, the class conditional probability $P(B_t^r|w)$ and prior probability $P(w)$ are first modeled and Bayes rule is then applied to infer the posterior probability $P(w|B_t^r)$. The examples of generative classifiers for SBD include [7], [9], [18]. While in discriminative classifiers, the posterior probability $P(w|B_t^r)$ is straightforwardly assumed in some functional form, and then the parameters of the function are estimated from the training data. The popular thresholding scheme is the simplest discriminative classifier, in which Ψ is assumed as the step function and the threshold is the unique parameter. Other applications of discriminative classifiers for SBD include [12]–[14]. The comparison of generative classifiers versus discriminative ones is an interesting topic of machine learning field, an in-depth discussion can refer to [19].

B. Major Challenges to the Formal Framework

To achieve satisfactory detection performance, special attention has to be paid to deal with several challenges to the above framework. Usually, the following three issues, i.e., the detection of GTs, the elimination of disturbances caused by abrupt illumination change or large object/camera movement, have been found the major challenges to current SBD techniques. How to conquer these challenges are the major difficulties while constructing the mappings in the proposed formal framework.

1) Detection of Gradual Transitions: As mentioned in Section I, although the detection of hard CUTs has been tackled, the detection of GTs remains a difficult problem. In [20], Lienhart presents an in-depth analysis on why the detection of GTs is more difficult than that of CUTs in the perspective of the temporal and spatial interrelation of the two adjacent shots. Here, from a different point of view, we summarize three reasons why it is difficult. First, GTs include various special editing effects, including dissolve, wipe, FOI, etc.. Each effect results in a distinct temporal pattern over the continuity signal curve. Second, GTs exhibit varying temporal duration, probably from three to dozens of frames. During a GT, although the continuity values of intra-frame features are usually smaller than those of within shots, they are not as significantly low as those of hard CUTs. Finally, the temporal patterns of GTs are similar to those caused by object/camera movement, since both of them are essentially processes of gradual visual content variation.

2) Disturbances of Abrupt Illumination Change: Most of the content representation methods are based on the color feature, in which luminance is a basic element. Abrupt illumination changes such as flashlights within shots often cause significant discontinuities of inter-frame feature, which is often mistaken for shot boundaries. Several illumination-invariant features and similarity metrics have been proposed to deal with the problem. However, these methods usually face a difficult dilemma, that is, illumination-invariant methods can certainly remove some disturbances of illumination change but they also lose the information of illumination change which is critical in characterizing the variation of visual content.

3) Disturbances of Large Object/Camera Movement: Besides shot transitions, object/camera movements also lead to the variations of visual content. Sometimes, the abrupt motion will cause similar continuity values to those of hard CUTs. Most of the times, the persistent slow motion will result in temporal patterns over continuity signal curve similar to those of GTs. It is difficult to distinguish the motion from the shot boundaries only using color features, since the behaviors of content variation are similar. The possible ways to handle the difficulties include adopting motion-compensated features or incorporating the features of motion activity.

III. Survey of the Existing Approaches

With the emergence of numerous SBD approaches, several excellent surveys have been presented [8], [9] [15], [21]–[23]. In this section, we do not attempt to present an exhaustive enumeration of the existing methods but focus on categorizing and analyzing them in the guide of the formal framework of Section II. Especially, some recent advances of SBD have been covered to complement the previous surveys. The methods discussed here will be categorized according to theirs roles in the formal framework. The pros and cons of various methods are identified by comparing the techniques of the same role, meanwhile, the optimal criteria of developing each separate module are discussed.

A. Methods of Visual Content Representation

There have been an intensive research on the representation approaches of visual content. Various techniques such as pixel-based [24]–[26], histogram [15], edge [27], motion [28], and even the mean and standard deviation of intensities [22] have been proposed. The comparison and evaluation of these methods are one of the focuses of previous surveys. In [8], [9], [15], [21], the performances of various approaches were evaluated. Different from other surveys, Lefèvre *et al.* concentrated on comparing the computational complexity of various approaches [23]. Several experimental evaluations have shown that the simple histogram feature usually is able to achieve a satisfactory result while some complicated features such as edge can not outperform the simple feature [15], [22]. In the following, we will concentrate on analyzing the tradeoff between the *invariance* and the *sensitivity* of various representation approaches.

The pixel-based method is the simplest method of constructing the mapping Φ, which maps each image to itself. Obviously, this is the most sensitive method, since it has captured any details of the frame. To speed the efficiency of pixel-based methods, several methods, known as visual rhythm [29], [30] or spatio-temporal slice [31], subsampled the pixels from the particular positions of each frame to represent the visual content. People have found that the pixel-based approach is somewhat sensitive to local or global movement. To handle the drawbacks, several variants of pixel-based method have been proposed. For example, Zhang *et al.* proposed to smooth the images by a 3×3 filter before performing the pixel comparison [26]. Color histogram, which captures the ratio of various color components or scales, is a popular alternative of the pixel-based methods. Since the color histogram does not incorporate the spatial distribution information of various colors, it is more invariant to local or small global movements than pixel-based methods. However, it is not expressive enough to distinguish the shots within the same scene. A better tradeoff between pixel and global color histogram methods can be achieved by block-matching methods, in which each frame is divided into several nonoverlapping blocks and the histogram feature or others of each block are extracted [32].

The aforementioned features mainly reflect the color intensities of visual content. Features describing the structural information of each frame are also proposed. For example, Zabih *et al.* proposed an edge change ratio (ECR) method to perform SBD [27]. In the so-called ECR method, Canny edge detector is employed to calculate the edge map of each frame, i.e., the structural representation of the visual content. Lienhart compared the ECR-based hard CUT detection against histogram based methods. The experiments reveal that ECR usually do not outperform the simple color histogram methods, but are computationally much more expensive [22]. Despite this depressing conclusion, the edge feature finds their applications in removing the false alarms caused by abrupt illumination change, since it is more invariant to various illumination changes than color histogram. Kim *et al.* [33] and Heng *et al.* [34] independently designed flashlight detectors based on the edge feature, in which edge extraction was required only for the candidates of shot boundaries and thus the computational cost was decreased.

Both the illumination and the structural layout are the important aspects of visual content. For the invariance, we prefer to the features invariant to illumination and structural layout, while for the sensitivity, we prefer to the ones capturing the variation of illumination and structural layout. It is difficult to develop a single approach which is not only invariant to various disturbances but also sensitive enough to capture the details of visual content. Integrating several complementary features to represent the visual content is probably a promising way.

B. Methods of Constructing Continuity Signal

Here, we do not consider the representation of content and the calculation of continuity together as what the previous surveys do. We single out the calculation of continuity as a separate problem. Furthermore, we classify the existing methods into two categories according to whether they have incorporated the contextual information, i.e., whether $d = 1$ or $d > 1$ in (2). Generally, in most of the previous approaches, $d = 1$ holds while in several recent methods $d > 1$ holds.

1) Pair-Wise Comparison Scheme With $d = 1$: The most straightforward way to evaluate the continuity of I_t and I_{t+1} is to directly compare their features

$$s_t = \Theta(V_t, V_{t+1}). \tag{5}$$

Concretely, the method of constructing Θ depends on what kinds of content representation method adopted. In pixel-based methods, Θ is obtained by comparing the corresponding pixels between I_t and I_{t+1}. With histogram methods, chi-square test and intersection have been tried to calculate Θ [15]. While in edge-based methods, the matching ratio of edge maps of the adjacent frames is used [27]. To obtain a motion independent metric, the mapping Θ can be constructed by block matching [9], where s_t is defined as the accumulation of the continuities between the most suited block-pairs of I_t and I_{t+1}.

One major drawback of the pair-wise comparison scheme is its sensitivity to noises. If a frame is distorted by noises, e.g., flashlight frame, the continuities between it and the two neighboring frames are usually dramatically small. It is often mistaken for a shot boundary. There exist several techniques refining the original continuity signal to suppress the disturbances of various noises. Leszczuk *et al.* [35] and Zheng *et al.* [36] proposed a so-called second-order difference method to construct the discontinuity signal. Their experiments show that the method can effectively reduce some disturbances of motion. In [37], Jun *et al.* proposed to first smooth the original signal by a median filter, and then subtract the smoothed one from the original signal, finally obtain a clear measured signal. Actually, these techniques of refining the signal are some implicit ways of using the contextual information of the nearby temporal interval.

2) Contextual Information Scheme With $d > 1$: Hanjalic pointed out that as much additional information as possible should be embedded into the shot boundary detector to effectively reduce the influence of the various disturbances [9]. For example, not only should the variation between the adjacent frames [i.e., $d = 1$ in (2)] be examined but also the variations within the temporal interval nearby [$d > 1$ in (2)] should be investigated, i.e., contextual information. Only recently have the methods explicitly using contextual information appeared [13], [38], [39]. Cooper Summarized these ideas as a similarity analysis framework to embed the contextual information [13]. First, a similarity matrix is generated by calculating the similarities between every pair of frames in the video sequence. Next, the frame-indexed score, i.e., the continuity signal, is computed by correlating a small kernel function along the main diagonal of the matrix. Designing an appropriate kernel function for correlation is the critical issue within this method. Cooper carried out a comparison of four kernel functions. However, the experimental results were not well fit to the intuitional assumption. In Section V, we will provide the physical interpretation for each kernel via graph partition model. Meanwhile, we will compare the approach proposed in Section IV with the kernel correlating methods. Note that the methods with $d > 1$ embed the contextual information while constructing the continuity signal, which is different from the pair-wise comparison scheme $(d = 1)$ which incorporates contextual information by additional post-processing procedure.

C. Methods of Classification

In Section II, we have categorized the existing classification methods into generative and discriminative methods. Here, to emphasize the evolution track of SBD, we prefer another type of taxonomy, namely rule-based methods and statistical machine learning ones.

1) Rule-Based Classifiers: In rule-based classifiers, the classification function is usually defined as:

$$w_t = \Psi\left(B_t^0\right) = \begin{cases} 0, & \text{if } s_t > T \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

where T is a predefined threshold. In the above equation $r = 0$ holds for B_t^r, which means only the amplitude of the continuity signal s_t is considered. If the continuity value exceeds the threshold T, the classifier outputs 0 to indicate no transition occurs between I_t and I_{t+1}, otherwise, the classifier yields 1 to declare the occurrence of shot transition. In the early work, heuristically chosen global thresholds were used. It is difficult to select a threshold T appropriate for various genres of videos. To address this drawback, various local adaptive thresholds were proposed [16], [40], [41]. The basic idea is to use a sliding window going along the continuity signal and computing the threshold locally within the sliding window. The local adaptive scheme incorporates the contextual information by taking the local activity of the content variations into account. The previous experiments showed its superiority over the global thresholding scheme. The related surveys with discussions on thresholding scheme can be found in [8], [9]. It should be notified that, in rule-based methods, the shapes of the classification hyperplane are actually manually designed, which requires the developers to be familiar with the characteristics of various genres of videos.

2) Statistical Machine Learning: There have been some recent efforts treating SBD as a pattern recognition problem and turning to the tools of machine learning. In Section II, the machine learning methods are divided into two categories, i.e., generative and discriminative classifiers. Generative classifier meets the requirements of explaining the generation mechanisms of shot transitions while discriminative classifier usually seems to be a black box. Furthermore, generative classifier is usually more convenient to incorporate additional information (e.g., *a prior* information). For example, in [7], [9], and [18], the shot duration was modeled and used to improve the detection performance, while all of the SBD system based on discriminative classifiers only used the feature of content variation activity. Nevertheless, generative methods usually highly depend on the assumptions of prior information and the models of class conditional distributions [19]. We have to make sure the correctness of the model assumptions before using generative classifiers. If the above requirements can not be satisfied, the discriminative classifier is preferred. Various discriminative approaches, including K-means [42], KNN [13], and support vector machines (SVMs) [12], [43]–[45], have been employed to perform SBD. With the statistical machine learning methods, the parameters of the models are chosen via cross validation processes and the shapes of classification hyperplane are constructed automatically during the training procedure. There are two key problems while utilizing machine learning methods. The first one is how to construct the features for the classifiers. Cooper [13] and Yuan *et al.* [12] used the continuity signals within the particular temporal interval as the features for KNN and SVMs, respectively. Similarly, Feng *et al.* [45] adopted the wavelet coefficient vectors within a sliding window as the features of SVMs. The second key problem is how to obtain a well-chosen training set with relatively balanced positive and negative examples, since within each video sequence the number of negative examples usually significantly exceeds that of positive examples. To address the problem, Lienhart [20] used a dissolve synthesizer to create an infinite amount of dissolve examples and produce the nondissolve pattern set by means of so called bootstrap method. Chua *et al.* [44] and Yuan *et al.* [12] adopted the active learning strategy to handle the imbalance training data. Compared with the thresholding schemes, the machine learning methods make decisions via the recognition of the shot transition patterns instead of the evaluation of the amplitude of content variations. It is expected that full use of contextual information can be made by machine learning methods.

D. Methods of Gradual Transition Detection

As mentioned in Section II, the detection of GTs is one of the major challenges to the proposed formal framework. So far, no techniques of GT detection have been able to achieve the result comparable to that of CUT detection. Some of the existing methods are designed to detect one specific editing effect, such as FOI, wipe and dissolve, while others are developed to detect several types of editing effects simultaneously. The relatively comprehensive surveys can refer to [8] and [9]. In the following, we present a brief overview of the existing methods for the sake of completeness.

1) Fade Out/In: During the FOI, two adjacent shots are spatially and temporally well separated by some monochrome frames [20], whereas monochrome frames seldom appear elsewhere. Lienhart proposed to first locate all monochrome frames as the candidates of FOIs [22]. Thus, the key of the FOI detection is the recognition of monochrome frames. For this purpose, the mean and the standard deviation of pixel intensities are commonly adopted to represent the visual content. The effectiveness of monochrome frame detection has been reported in [8], [36], and [40].

2) Wipe: For wipes, the adjacent shots are not temporally separated but spatially well separated at any time [20]. An interesting method for wipe detection is the so-called spatio-temporal slice analysis [31]. For various styles of wipes, there are corresponding patterns on the spatio-temporal slices. Based on this observation, Ngo *et al.* transformed the detection of wipes to the recognition of the specific patterns on spatio-temporal slices. Other wipe detection methods such as [46] are also based on the fact that two adjacent shots before and after wipes are spatially well separated at any time.

3) Dissolve: In the process of dissolve, two adjacent shots are temporally as well as spatially intermingled [20]. Hampapur *et al.* [47] proposed an approach based on the production model of dissolve, which highly depends on the definition of the chromatic scaling functions. Since the durations and mixing styles of different dissolves vary abroad, it is difficult to define a single scaling function suitable for all the dissolves. Furthermore, the assumption that no motion exists during the dissolve procedure is usually not satisfied. In the result, the actual performance of the model-driven methods is not satisfactory. Another popular dissolve detection method is based on the characteristic of the change of intensities variance, i.e., the so-called downwards-parabolic pattern, which was originally proposed by Alattar [48]. In the result of thorough mixture of the two adjacent shots, the variance of the pixel intensities will decrease from the beginning of the dissolve and reach the minimum in the middle of the transition. The detection of dissolve becomes the recognition of the parabolic pattern on the variance curve. Several improvements on this idea can be found in [31] and [40]. The relatively satisfactory results have been reported in [8].

4) General Approaches for Gradual Transitions: With global color feature adopted, various types of GTs such as wipes and dissolves exhibit similar characteristics over the continuity signal curve. Therefore, it is possible to develop a unified technique to detect several types of GTs simultaneously. For example, the well-known twin-comparison technique, proposed by Zhang *et al.* [49], is a general approach to detect GTs. Nevertheless, it often truncates the long GTs because of the mechanism of the global low threshold. In addition, it has difficulties in reducing the disturbances of motion. To overcome the shortcomings, Zheng *et al.* [36] proposed an enhanced twin-comparison method, i.e., finite state automata (FSA) method, in which motion-based adaptive threshold was utilized. This method yielded the best performance of GT detection on the benchmark of TRECVID 2004. Different from CUTs, GTs span varying temporal duration, which makes it difficult for a single fixed scale transition detector to detect all the GTs. The success of the twin-comparison based methods is somewhat due to the exploitation of the multiresolution property of GTs, i.e., low threshold for high resolution and high threshold for low resolution. Several other methods have been proposed in the form of explicit temporal multiresolution analysis. Lin *et al.* [50] and Chua *et al.* [44] exploited the multiresolution edge phenomenon in the feature space and designed a temporal multiresolution analysis (TMRA) based algorithm which used Canny wavelets to perform temporal video segmentation. The experimental results showed that the method could locates CUTs and GTs in a unified framework. However, as noted by the author, the Canny wavelet transform is computationally intensive. Another multiresolution idea is to adjust the sample rate of the video. For example, Lienhart [20] employed a fixed scale transition detector to run on sequences of different resolutions to detect dissolves. Similarly, Ngo [43] reduced the problem of dissolve detection to a CUT detection problem in a multiresolution representation. In Section IV, we will introduce another GT detection method based on the temporal multiresolution analysis of graph partition model [12].

IV. SBD System Based on Graph Partition Model

In this section, we will introduce a unified SBD approach based on graph partition model. The shorter versions of this work have been separately published in [12], [51], and [52]. Here, we do not intend to redescribe it but focus on demonstrating how the criteria discussed in the above sections are carried out in an effective framework. Since the previous work has shown that color histogram is a suitable representation method of visual content, here we directly adopt it to represent the content of each frame. This section will focus on dealing with the remainder three problems, i.e., the construction of continuity signal, the classification of continuity values and the detection of GTs.

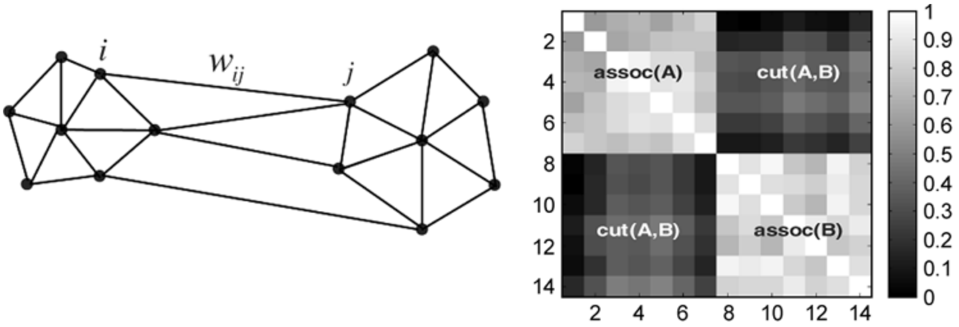


Fig. 1. Left: A Graph with 14 nodes. Right: Visualization of the similarity matrix of the left graph. w_{ij} is defined as the reciprocal of Euclidean distance of the nodes i and j. The stronger the connectivity between i and j, the brighter the entry (i, j) is.

A. *Graph Partition Model*

In [51], we have proposed a graph partition model to perform temporal data segmentation. Video is a typical kind of high dimensional temporal data, thus in [12] this model is used to perform SBD. Before presenting the SBD framework, we first introduce some prerequisite knowledge on graph partition model, and then show how it is applied to fulfill temporal video segmentation.

1) Segmentation by Graph Cuts: Given an undirected, weighted graph $G = G(V, E)$ with a set of nodes V, a set of edges E. Assume $|V| = N$, namely there are N nodes in graph G. Let $w_{ij} \in [0, 1]$ denote the weight of edge $e(i, j) \in E$, which indicates the similarity between nodes i and j. The larger the w_{ij}, the more similar between nodes i and j. To introduce the graph partition model more clearly, we first define several graph terminologies:

Definition 1: The *similarity matrix* $\mathbf{W}$ is a $N \times N$ symmetric matrix, in which entry w_{ij} indicates the similarity of nodes i and j.

Definition 2: The cut[1] which divides graph G into subgraphs A and B is defined as: $\mathrm{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}$.

Definition 3: The *association* of subgraph A is defined as: $\mathrm{assoc}(A) = \sum_{i,j \in A} w_{ij}$.

An example of graph and the related terminologies are illustrated in Fig. 1. As shown in the example, the $\mathrm{cut}(A, B)$ reflects the strength of the connectivity between the two subgraphs A and B, while the $\mathrm{assoc}(A)$ and $\mathrm{assoc}(B)$ reveal the strength of the connectivity within the subgraph A and B, respectively.

Given a data set, a graph can be constructed by treating each sample as a node and linking an edge between each pair of nodes. By defining the weight of edge as the similarity of samples, data segmentation can be formulated as a graph partition problem. Various partition criteria can be defined. Initially, minimum cut is proposed to be a partition objective function. However, it usually leads to the skewed cut. Therefore, several other objectives such as ratio cut [53], normalized cut [54] and min-max cut [55] are proposed successively. From the point of view of segmentation, min-max cut defined by (7), which tries to minimize the association between the two subgraphs while maximize the association within each subgraph, gives the best criterion

$$M\mathrm{cut}(A, B) = \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(A)} + \frac{\mathrm{cut}(A, B)}{\mathrm{assoc}(B)}. \tag{7}$$

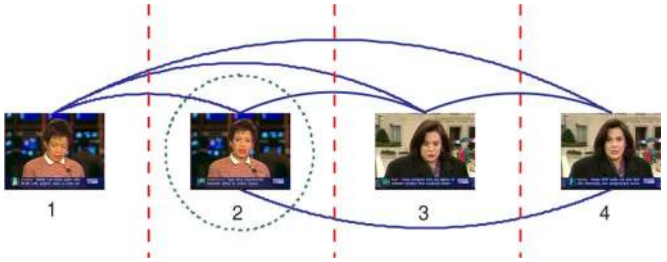


Fig. 2. Full graph constructed from four successive frames. The solid lines represent the edges between the frames, and the dash lines indicate the positions of the feasible cuts while the dot circle is a nonfeasible cut. For the full graph with four nodes, there are ten possible cuts. However, to segment the video sequence to some shots, there are only three of them reasonable.

It is expected that the global minimum solution of $M\mathrm{cut}(A, B)$ will yield the optimal partition. Unfortunately, the problem is NP-complete because of its combinatory nature [55]. A popular approach, namely spectral graph partition, has been proposed to get an approximate optimal solution, which is based on the spectral graph theory [54], [55]. Still it can not handle a huge amount of data because of the intensive computation while performing the matrix spectral decomposition.

2) Cuts With Temporal Constraints: When applied to the problem of SBD, the graph partition model must satisfy some temporal constraints. For example, the approach must guarantee the temporal continuity of each shot. In other words, once two un-adjacent frames are grouped into the same shot, any frame between them must be clustered into the same one. Imposing this temporal constraint on the model, a feasible partition can only occur at one of the $N - 1$ possible positions between any two adjacent frames. Thus, as shown in Fig. 2, the size of feasible set is reduced from exponential to $N - 1$. To get the optimal solution, we just need to compute the objective values of the $N - 1$ feasible cuts, and then select the one with minimum objective via a linear search. Formally, we define $\mathrm{score}(t)$ as the objective function of the tth feasible cut

$$\mathrm{score}(t) = M\mathrm{cut}\left(\{1, 2, \ldots, t\}, \{t + 1, t + 2, \ldots, N\}\right). \tag{8}$$

[1] To avoid confusion, we mean the abrupt shot transition by the capital word "CUT," and indicate the terminology of the graph theory by the lowercase word "cut."

Then the cut with minimal score is the optimal solution. By imposing the constraints of temporal continuity, the time complexity of the partition problem has been dramatically decreased. The above procedure can only partition a graph into two subgraphs. To segment data into more than two segments, we can partition the subgraph recursively. Or alternatively, when the score of a cut is sufficiently small, a boundary is declared there.

3) Temporal Video Segmentation: Given a video sequence, by treating each frame as a node and linking each other with an edge, we can construct a weighted graph $G(V, E)$. In this way, the SBD is formulated as a graph partition problem.

a) How to define w_{ij}?: The weight w_{ij} of the edge between the ith and the jth frame should reflect the likelihood that the two frames belong to the same shot. The more similar the frames I_i and I_j are, the higher w_{ij} should be. On the other hand, the larger the temporal distance between them, the lower the probability that they belong to the same shot. In the extreme, if a frame is too far from the other, it is impossible for them to be in the same shot, that is, $w_{ij} = 0$. Let H^i be a k bins color histogram of the ith frame and adopt histogram intersection method to measure the similarity, a reasonable definition of w_{ij} is

$$w_{ij} = \sum_k \min\left(H_k^i, H_k^j\right) \times \begin{cases} e^{\frac{-\|i-j\|_2^2}{\sigma^2}}, & \text{if } |i-j| < d \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where σ is a factor reflecting the similarity decaying with the temporal interval increasing, and d indicates the same meaning with that of (2). We can find that the above definition of w_{ij} essentially equals to imposing a gaussian filtering along the main diagonal of the similarity matrix. With the restriction of d, the calculation of $\mathrm{score}(t)$ is restricted in a $4 \times d^2$ submatrix, which we call *active matrix*. Therefore, the (8) can be simplified as

$$\mathrm{score}(t) = M\mathrm{cut}\left(\{t-d+1, \ldots, t\}, \{t+1, \ldots, t+d\}\right). \quad (10)$$

In the end, the construction method of the continuity signal is depicted by the following equation:

$$s_t = \Theta\left(A_t^d\right) = \mathrm{score}(t). \quad (11)$$

Note that the value of $\mathrm{score}(t)$ is determined by the content variation within a interval of range $2 \times d$. Therefore, for a CUT transition between I_t and I_{t+1}, the corresponding temporal pattern over the curve of continuity signal is not an isolated local minimum but a valley shape local minimum, in which the scores from s_{t-d+1} to s_t gradually decrease to the minimum and from s_t to s_{t+d} the score will gradually increases to the normal values. As shown in Fig. 3, there is a sharp valley corresponding to each CUT. The identification of CUTs can be easily handled by the recognition of valleys.

b) The Algorithm: In summary, if the thresholding scheme is adopted to detect CUTs, the video temporal segmentation algorithm consists of the following steps.

Step 1) Given a video sequence, treat each frame as a node and link each other by an edge, to construct a weighted graph $G(V, E)$.

Step 2) Compute w_{ij} according to (9), obtaining a similarity matrix $\mathbf{W}$.

Step 3) Calculate scores of the $N-1$ feasible cuts according to (10).

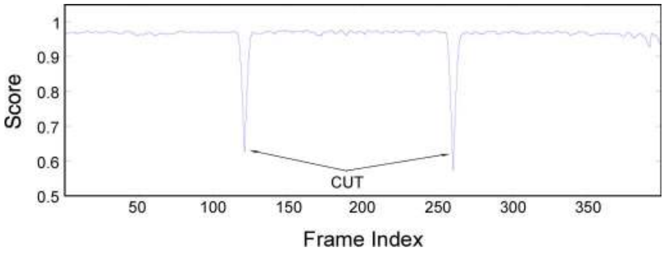


Fig. 3. Segment of continuity signal calculated according to (10), in which $\sigma = 150$ and $d = 5$.

Step 4) Select feasible cuts whose scores are the local minima of the corresponding neighborhoods within a radius of d.

Step 5) Declare the cuts whose scores are below a pre-defined threshold as CUTs.

c) Analysis of the Algorithm: With the definition of *active matrix* in size of $4 \times d^2$, it is not necessary to involve all the entries while computing each s_t. With a video sequence of length N, the overall time complexity is $O(N \times d)$ with the overlapping of two successive active matrices considered. Compared with the spectral graph partition methods, it is much more efficient.

Another prominent advantage of the approach is the invariance. Instead of pair-wise comparison with $d = 1$, the graph partition method performs boundary detection by considering the feature variations in a local neighborhood, i.e., the contextual information with $d > 1$. As shown in Fig. 4, there are three flashlights occurring within the video sequence. In the pair-wise comparison method with $d = 1$, the three corresponding sharp valleys are usually classified as shot boundaries because of the low amplitudes comparable to those of CUTs. In the result, a lot of false alarms are caused. While in the proposed approach, the strong connectivity between the frames before and those after the flashlight makes it unlikely to separate the sequence to two parts. There is no distinct undulation on the continuity signal, and therefore no specific flashlight detector is needed. Besides the disturbances of abrupt illumination change, the approach is also invariant to other abrupt noises. The experiment in Section V will further confirm this observation.

B. Support Vector Machine Active Learning

Having obtained the curve of the continuity signal, the shot boundaries can be identified by thresholding scheme as what most of the existing methods do. However, the thresholding method has several difficulties in achieving satisfactory results. First, the chosen threshold usually highly depends on the genres of videos. The intensities of some content variations caused by noises have exceeded those of shot transitions. In the result, any chosen global threshold can not distinguish the boundaries and nonboundaries successfully. Second, a single threshold can not make full use of the contextual information, such as whether the signal variation is a sharp valley or a gentle concave. This is significantly important to classify CUTs from GTs, or GTs from scenes with motion. Third, even in the adaptive thresholding scheme, some parameters are required to be heuristically

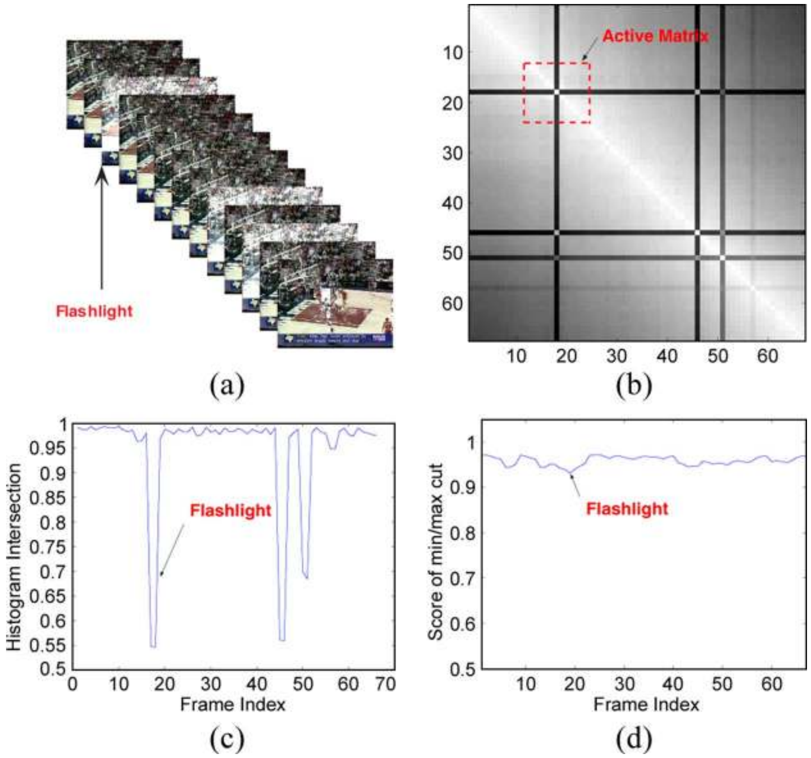


Fig. 4. (a)Video sequence with three flashlights occurring. (b) Visualization of the corresponding similarity matrix, in which the dash rectangle indicates the range of *active matrix*. (c) Continuity signal obtained by the comparison between the successive frames. (d) Continuity signal obtained according to (10).

determined, which still depend on the genres of videos. Therefore, the thresholding scheme is essentially a procedure of manually constructing the classification hyperplane, it requires the developers to be familiar with the characteristic of videos. In the following, we will introduce how to apply machine learning methods to address the drawbacks of thresholding scheme. To simplify the problem, we will focus on the CUT detection here, the method of GT detection will be presented in the next subsection.

1) Feature Construction: With the continuity signal produced by graph partition model, each CUT boundary corresponds to a local minimum on the curve. However, not every local minimum is a CUT boundary. Some of the local minimum are caused by GTs, and others result from the various disturbances. Only by evaluating the magnitudes of local minima can not successfully distinguish boundaries and nonboundaries. By observing the signal curves, we find that the valleys corresponding to the boundaries, especially for CUTs, are usually regular and somewhat symmetric, while valleys caused by disturbances are not. That is, the boundaries and nonboundaries can be better classified by recognizing the shapes of corresponding valleys. Formally, the shape of the valley centering at s_t can be characterized by the feature vector B_t^r

$$B_t^r = (s_{t-r}, \ldots, s_t, s_{t+1}, \ldots, s_{t+r}) \tag{12}$$

where r usually equals the variable d of (11), since a CUT boundary between I_t and I_{t+1} affects the signal values at most to s_{t-d} before s_t and at most to s_{t+d} after s_t.

2) Support Vector Machine: As for the selection of classifiers, SVMs is preferred, not only for its solid theoretical foundations but also for its various empirical success [56]. SVMs is an approximate implementation of the Structural Risk Minimization (SRM) induction principle. The main idea behind SVMs is to separate classes with a surface with maximal margin between them so as to minimize the risk of over-fitting. Thus, SVMs is not only with simple geometric explanation but also with an elegant formulation as a quadratic optimization problem. Because of the convexity of the quadratic problem with linear and box constraints, the global optimum solution can be guaranteed. There are two attractive properties for the optimum solution in dual representation. First, only the support vectors appear in the solution. Second, the support vectors only appear in the form of dot products, which makes the kernel trick possible to handle the curse of dimensionality. The final decision function output by SVMs is in the following form:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + b\right) \tag{13}$$

in which the symbols are of the same definitions with [56]: $\mathbf{x}$ and $\mathbf{s}_i$ represent the input vector and support vector, respectively; N_s indicates the number of support vectors; the α_i and y_i are the Lagrange multiplier and class label corresponding to $\mathbf{s}_i$, respectively; K is the kernel function. Concretely in our problem, the mapping from the continuity value space to the decision space is defined as

$$w_t = \Psi\left(B_t^r\right) = \operatorname{sgn}\left(\sum_{i=1}^{N_s} \alpha_i y_i K\left(\mathbf{s}_i, B_t^r\right) + b\right). \tag{14}$$

3) Active Learning Strategy: To train an SVMs model, we have manually annotate a training set consisting of positive ex-

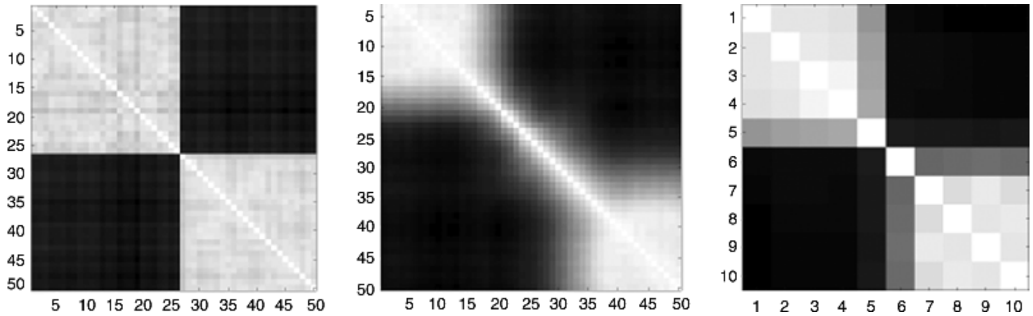


Fig. 5. Left: Pattern of a CUT boundary on the similarity matrix. Middle: Pattern of a GT boundary on the similarity matrix.Right: View of the middle pattern at a lower resolution.

amples and negative examples. In the annotation procedure, we usually explicitly identify the positive examples while implicitly consider all the other intervals as negative examples. This leads to an imbalanced training set consisting of limited positive examples and infinite negative examples. It is impossible to directly train SVMs model on this raw training set. First, the training time is unacceptable with the large training set. Second, SVMs usually can not converge to a reasonable model with the severely imbalanced training set. To speed up the training procedure of SVMs, Schohn *et al.* [57] proposed an active learning method with SVMs. The basic idea is based on the sparseness of the solution to SVMs model. As shown in (14), the decision function is determined by the support vectors, and the other training examples which are far from the hyperplane do not have influence on the position and shape of the decision function. Therefore, Schohn proposed to select the examples lying closest to the SVMs' dividing hyperplane as training set. The experiments show that the method offers better performance with fewer training data.

Here we adopt a heuristic active learning strategy to handle the imbalanced training set. According to the feature construction method, CUTs are distinguished from the others via the shapes of valleys. We have observed that all the intervals corresponding to the shot boundaries are in the shapes of valleys, but not all of the valleys are caused by shot transitions. Therefore, the intervals in valley-shape are assumed to lie closest to the dividing hyperplane and should be selected as training examples. Formally, the requirement for the selected signal intervals is

$$\begin{cases} s_t = \min\{s_{t-r}, \ldots, s_t, s_{t+1}, \ldots, s_{t+r}\} \\ s_t \leq T_{\mathrm{CUT}} \end{cases} \tag{15}$$

where T_{CUT} is a pre-defined threshold that guarantees the local minimum s_t is sufficiently small. All the intervals satisfying (15) are collected. The real boundaries are labeled as positive examples, and the others are labeled as negative examples. To achieve the utmost performance, the threshold T_{CUT} should be selected carefully. The larger T_{CUT} is, the more negative examples are in the training set. If T_{CUT} is too large, the number of negative examples will significantly exceed that of positive ones. Inversely, if T_{CUT} is too small, many of positive examples will be removed from the training set. A suitable method of selecting T_{CUT} is cross validation.

C. *Temporal Multiresolution Analysis*

In the previous sections, we have introduced how the graph partition model and the SVMs can be applied to detect CUTs. GTs exhibit distinct characteristics from CUTs. Several specific techniques have to be designed to handle the challenges of GTs.

1) The Problem: For CUT, the transition occurs between two adjacent frames belonging to the first shot and the second one, respectively. Due to the characteristic of abrupt content variation, there is a clear correspondent "chessboard" pattern on the similarity matrix, as shown in Fig. 5. Thus, with single values of d and r, it is sufficient to detect all of CUTs. Different from CUTs, GTs are more difficult to detect. On the one hand, GTs may span a varying temporal length. It is difficult to cover all the situations with single d and r. On the other hand, the content variation between adjacent frames may be rather small, yielding a blurry pattern on the similarity matrix, as shown in Fig. 5. In the result, the content continuity signal does not vary significantly enough to reflect the existence of shot transitions. The above characteristics of GTs determine that we have to design distinct active learning, feature construction and classification methods from those of CUTs.

2) Temporal Multiple Resolution Analysis: As shown in Fig. 5, although GT is not observable at a high resolution, it can be easily detected at a lower resolution. This phenomenon has been observed and exploited in the form of multiresolution analysis in [20], [43], [44], [50]. The objective of multiresolution analysis is two-fold. First, with the multiresolution analysis, it is possible to detect GTs of various durations with the feature vectors of fixed length, which will facilitate the training of SVMs. Second, with the fusion of multiresolution results, it is expected to boost the detection performance. Generally speaking, the multiresolution analysis can be performed at every phase of the formal framework. For example, the most straightforward way is to adjust the sampling rate of the video sequence [20], [43]. Alternatively, it is can also be performed by constructing the continuity signals in multiple scales [50]. Finally, multiresolution can also be carried out by constructing feature vectors of multiple scales from a single continuity signal. In the following, we will introduce two schemes corresponding to the latter two levels.

a) Construction of multiresolution graph: In this method, with the variation of the sampling rate of video sequences, multiple continuity signals at several resolutions are calculated. Let

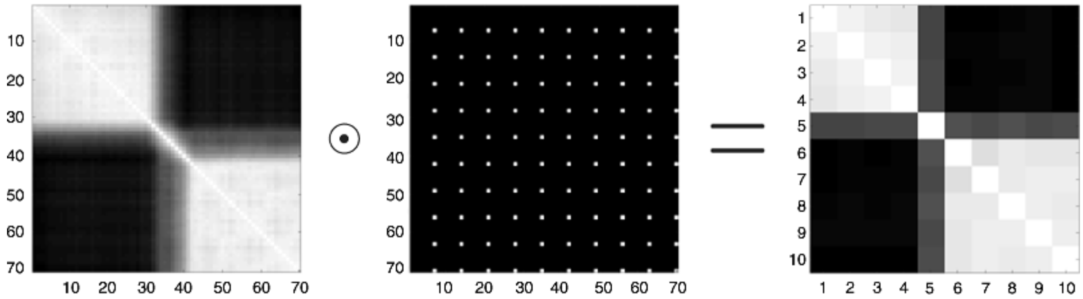


Fig. 6. Three matrices are $\mathbf{W}_t^{\delta}$, $\mathbf{M}^{\delta}$, $\mathbf{V}_t^{d}$, respectively, where $d = 5$ and $\delta = 7$. The operation "$\odot$" means the Hadamard multiplication. The white blocks in $\mathbf{M}^{\delta}$ represent entries 1. The first two matrices are square ones of width 70, while the last one is square matrix of width 10.

$\delta \in \{1, 2, \ldots\}$ be the sampling rate of frames, the score of the tth feasible partition at the δth resolution is defined as follows:

$$\begin{aligned}\mathrm{score}(t,\delta) = M\mathrm{cut}\,(\{t-(d-1)\times\delta,\ldots,t-\delta,i\}\,,\\ \{t+\delta,\ldots,t+d\times\delta\})\,. \quad (16)\end{aligned}$$

The above equation shows that the algorithm samples every δ frames in a larger range of $\{i-(d-1)\times\delta,\ldots,i+d\times\delta\}$ instead of involving all the frames in a neighborhood of $\{t-d+1,\ldots,i+d\}$. With the δ varying, multiple temporal resolution graphs can be constructed. Let s_t^{δ} denote $\mathrm{score}(t,\delta)$ defined in (16), the shape of the valley centering at t over the δth resolution signal can be characterized by the feature vector B_t^{δ}

$$B_t^{\delta} = \left(s^{\delta}_{t-r\times\delta}, s^{\delta}_{t-(r-1)\times\delta}, \ldots, s^{\delta}_{t}, s^{\delta}_{t+\delta}, \ldots, s^{\delta}_{t+r\times\delta}\right). \quad (17)$$

Note that for different δ, the feature vector B_t^{δ} is with the same length, i.e., $2\times r$. While implementing the above algorithm, we define a square *selective matrix* $\mathbf{M}^{\delta}$ of width $2\times d\times\delta$, whose entries can only equal 0 and 1, 0 indicating the the corresponding frame is not sampled, 1 representing the frame sampled. Let $\mathbf{W}_t^{\delta}$ be the square submatrix of $\mathbf{W}$, centering at t and with the width $2\times d\times\delta$. To calculate $\mathrm{score}(t)$ at the resolution δ, the algorithm performs Hadamard multiplication of $\mathbf{W}_t^{\delta}$ and $\mathbf{M}^{\delta}$, and results in a new matrix $\mathbf{V}_t^{\delta}$, in which the zero entries correspond to the nonsampled frames. Ignoring the 0 entries, $\mathbf{V}_t^{\delta}$ can be restricted to an equivalent but smaller square matrix $\mathbf{V}_t^{d}$ of width $2\times d$. The above process is depicted in Fig. 6. Via this operation, the temporal multiresolution analysis problem is transformed to the spatial multiresolution analysis on the similarity matrix.

b) Construction of multiresolution score: In the above approach, for each δ, a continuity signal is obtained. For each GT candidate, there is a corresponding feature vector at the curve of each resolution. However, the adjustment of sampling rate of frame sequences is a double-edged sword. With the lower sampling rate, the variations of continuity signal corresponding to the gradual transitions are made prominent. In the result, the GTs are easier to be observed. However, the variations caused by motion are also enlarged and more difficult to distinguish from GTs. Without effective ways to reduce the disturbances of motion, the above multiresolution analysis approach will not yield satisfactory results. An alternative is to keep the sampling rate of frame sequences unchanged but vary the sampling rate of the continuity signal to construct multiresolution feature vectors. Here, let δ be the sampling rate of continuity signal, the feature vector centering at t of the δth resolution is defined as

$$B_t^{\delta} = \left(s_{t-r\times\delta}, s_{t-(r-1)\times\delta}, \ldots, s_{t}, s_{t+\delta}, \ldots, s_{t+r\times\delta}\right). \quad (18)$$

3) Fusion of Multiresolution Analysis: The effectiveness of the information fusion has been demonstrated in various applications especially in content analysis of multimedia [58], [59]. Generally, there are two fusion schemes, namely early fusion and late fusion. It is difficult to prove one is superior to the other with theoretical analysis. However, it is possible to evaluate and compare them with experiments. For example, Snoek *et al.* compared them in the semantic video analysis [59]. Similarly, here we present the two fusion schemes and will evaluate them in Section V. Based on the statistical analysis of several videos, we find that almost all of GTs span the lengths from 3 to 100 frames. Therefore, with $\delta \in \{1, 3, 5\}$ and $r = 10$, the feature vector will spread long enough to cover most of GTs.

a) Early fusion strategy: In the early fusion, the feature vectors characterizing the same candidate (with the same t) at different resolutions (with different δ) are concatenated and form a single feature vector. Since only one kind of feature, only one learning phase is required and a unique SVMs model is obtained. Formally, the candidate centering at t is described by the following multiresolution representation:

$$B_t = \left(B_t^1, B_t^3, B_t^5\right). \quad (19)$$

b) Late fusion strategy: In contrast to the early fusion, where features are combined into a multiresolution representation and a unique model is trained, late fusion learns a model for each resolution. Then the outputs of the separate models are combined into the final decision, which can be achieved by various ways. In this paper, the final decision is obtained by the "OR" voting fusion of the outputs of the models at different resolutions. That is to say, once one model considers a candidate as a GT, the candidate will be declared as a true GT. Compared to the early fusion, the late fusion requires more learning processes.

4) Active Learning for Gradual Transitions: The implementation of the active learning idea for GTs is not as simple as that for CUTs as shown in (15). Here, the temporal intervals within which all the score values are below T_{GT} are first located

$$\begin{cases} s_{\mathrm{begin}} = s_{\mathrm{end}} = \max\{s_{\mathrm{begin}}, s_{\mathrm{begin}+1}, \ldots, s_{\mathrm{end}-1}, s_{\mathrm{end}}\} \\ s_{\mathrm{begin}} \le T_{\mathrm{GT}} \end{cases} \quad (20)$$

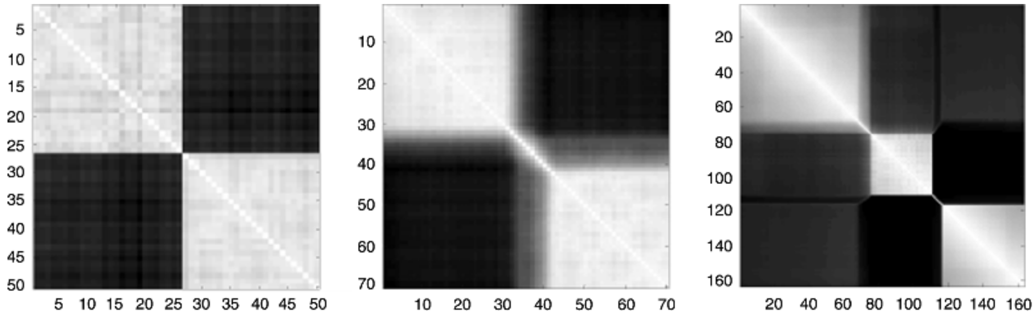


Fig. 7. From left to right: examples of patterns for cut, dissolve, and FOI on the similarity matrix.

in which $T_{\rm GT}$ is a pre-defined threshold with the similar role to $T_{\rm CUT}$ of (15). Then the local minimum of each selected interval is determined

$$s_t = \min\{s_{\rm begin}, s_{\rm begin+1}, \ldots, s_{\rm end-1}, s_{\rm end}\}. \tag{21}$$

Finally, centering at local minimum s_t, the multiresolution feature vectors are constructed according to (17) or (18). If the classifiers consider the feature vectors centering at s_t a GT, the $s_{\rm begin}$ and $s_{\rm end}$ will be declared as the start and the end frames of this GT.

D. Overview of the Whole System

Until here, we have to clarify that the above framework can not handle the detection of FOIs. During the process of FOIs, the first shot fades out and turns into a sequence of monochrome frames and then gradually the next shot fades in. As shown in Fig. 7, the FOIs patterns on the similarity matrix are different from those of CUTs and the other types of GTs. For CUTs and dissolves, there are two segments with coherent color feature before and after the shot transitions, while for FOIs there are three segments with coherent color feature, i.e., besides the two adjacent shots, an additional segment of monochrome frames between them. In the result, there are usually two "valleys" corresponding to an FOI. If we adopt the same detection approaches to those of CUTs and GTs, each FOI is usually classified as two shot transitions. Therefore, before applying the graph partition framework, specific technique is required to detect FOIs. In our implementation, an FOI detector based on the monochrome frame recognition is adopted [52]. To demonstrate the roles of each modules in the whole system, we will present a brief introduction to the system architecture. As shown in Fig. 8, the SBD is conducted by a hierarchical classification architecture. First of all, an FOI detector is employed to recognize the FOIs. Second, feature vectors for CUTs are constructed based on the graph partition model, and then are used to train an SVMs model or to be classified as CUTs and non-CUTs with the trained model. With all the FOIs and CUTs detected, multiresolution feature vectors are constructed to detect GTs. With the hierarchical classification procedure, all types of shot boundaries can be detected.

V. Experiments

In this section, we will carry out several comparison experiments on the platform of TRECVID. The first four are designed to evaluate the separate functional component of the proposed framework. The criteria discussed in the previous sections are verified. The last experiment is used to evaluate the whole proposed SBD system.

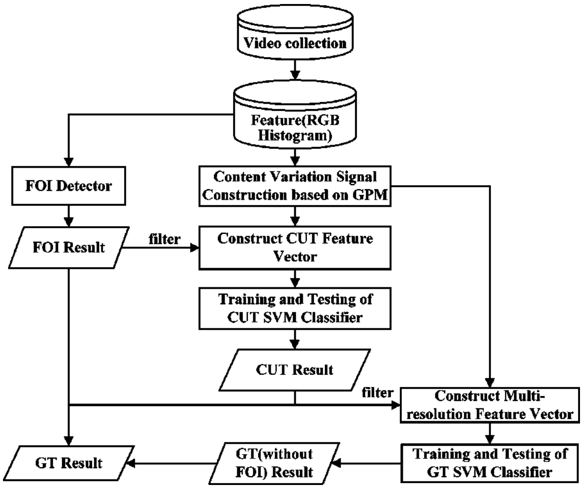

Video collection
Feature(RGB Histogram)
FOI Detector
Content Variation Signal Construction based on GPM
FOI Result
filter
Construct CUT Feature Vector
Training and Testing of CUT SVM Classifier
CUT Result
filter
Construct Multi-resolution Feature Vector
GT Result
GT(without FOI) Result
Training and Testing of GT SVM Classifier

Fig. 8. Flowchart of the proposed SBD system.

A. Experimental Setup

Both the annotated video collections and evaluation tool are provided by TRECVID. For the convenience of experiments, the video collections and evaluation criteria are slightly different from the original settings of TRECVID.

1) Data Corpora: All the 2003, 2004 and 2005 TRECVID test collections for SBD task are adopted. Totally, there are 31 videos in MPEG-1 format. For the details of each video, please refer to the homepage of TRECVID [6]. The summary of the collections of each year is listed in Table I. The original test collections for each year are labeled as "**D2003**," "**D2004**," and "**D2005**," respectively. Since the detections of CUTs and GTs interact, to reduce the factors that should be considered, some of the experiments are carried out to fulfil CUTs detection only. This helps to obtain accurate conclusions. For this purpose, we create another video collections by re-editing all the GTs to CUTs. Those collections are called "**D2003_NO_GT**," "**D2004_NO_GT**," and "**D2005_NO_GT**," respectively.

2) Evaluation Criteria: The output of the detection system is in XML format, which will be evaluated by the tool provided by TRECVID. Similar to other information retrieval task, the performance is evaluated by *recall* and *precision* criteria, which represent the fraction of relevant documents retrieved and the fraction of retrieved documents that are relevant, respectively.

TABLE I
SUMMARY OF THE TEST COLLECTION FOR SBD TASK IN TRECVID 2003–2005

Year	Number of Videos	Total Size (In gigabytes)	Total Frames	Total Transitions	Transition Types (Number/Ratio)				Sources
					Cut	Dissolve	FOI	Other	
2003	8	3.05	596054	3734	2644(70.7%)	753(20.2%)	116(3.1%)	221(5.9%)	ABC(4), CNN(4)
2004	12	4.23	618409	4806	2774(57.7%)	1525(31.7%)	230(4.8%)	276(5.7%)	ABC(6), CNN(6)
2005	12	4.65	744604	4535	2759(60.8%)	1382(30.5%)	81(1.8%)	313(6.9%)	CCTV(1), NBC(4), NASA(4), LBC(1), NTDTV(2)

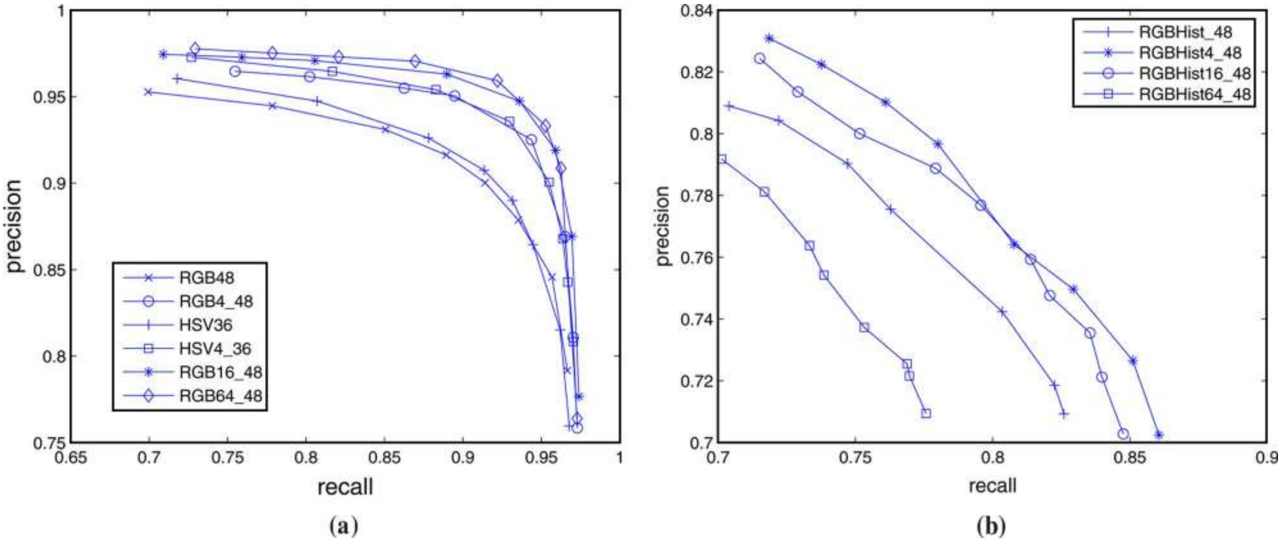


Fig. 9. Evaluation of content representation methods. (a) Performance of CUTs detection using various features. (b) Performance of GTs detection using various features.

For each approach, one or several parameters are varied and evaluated to obtain a curve of *precision* versus *recall*. The performances of different algorithms are compared via the *precision* versus *recall* curves. Sometimes, to rank performance of different algorithms, F_1 measure, a harmonic average of *recall* and *precision* is used. F_1 measure combining *recall* and *precision* with equal weight is in the following form:

$$F_1(\text{recall}, \text{precision}) = \frac{2 \times \text{recall} \times \text{precision}}{recall + precision}. \qquad (22)$$

B. Performance Comparison

Two types of experiments are carried out, namely module evaluation and system evaluation. For module evaluation, the specific module varies in different approaches while all the other modules of the system retain the same implementation. In this way, the differences between the compared systems are guaranteed to be caused by the difference of the evaluated module. For system evaluation, the proposed system is compared to those of the other participants of TRECVID 2005.

1) Visual Content Representation: As for the method of visual content representation, it has been thoroughly discussed and evaluated in the previous work [15]. Here, we focus on the evaluation of tradeoff between invariance and sensitivity. For this purpose, global color histogram and block-based color histogram is compared. In the block-based histogram, each frame is first divided into several blocks and the color histogram feature is extracted block by block. Totally, six kinds of feature are extracted.

RGB48	in RGB space, 16 bins for each channel
RGB4_48	2 × 2 blocks based **RGB48**
HSV36	in HSV space, un-uniform quantization [60]
HSV4_36	2 × 2 blocks based **HSV36**
RGB16_48	4 × 4 blocks based **RGB48**
RGB64_48	8 × 8 blocks based **RGB48**

With the above content representation methods, the performances of CUTs detection and GTs detection are evaluated, respectively, as shown in Fig. 9. For CUTs detection, the algorithm is evaluated on both the "**D2003_NO_GT**" and "**D2004_NO_GT**" collections. The graph partition model is used to construct continuity signal and thresholding scheme to make decisions. In the implementation, a global threshold is directly compared with the score curve and if the valley is below the threshold, a CUT is declared. By varying the threshold, the *precision* versus *recall* curve is obtained, as depicted in the (a)

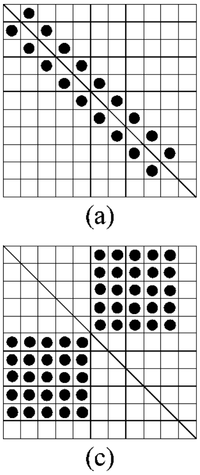


Fig. 10. Different kernels for SBD via kernel correlation $(\mathrm{r} = 5)$. (a) Scale space analysis. (b) Diagonal cross similarity. (c) Cross similarity. (d) Full analysis.

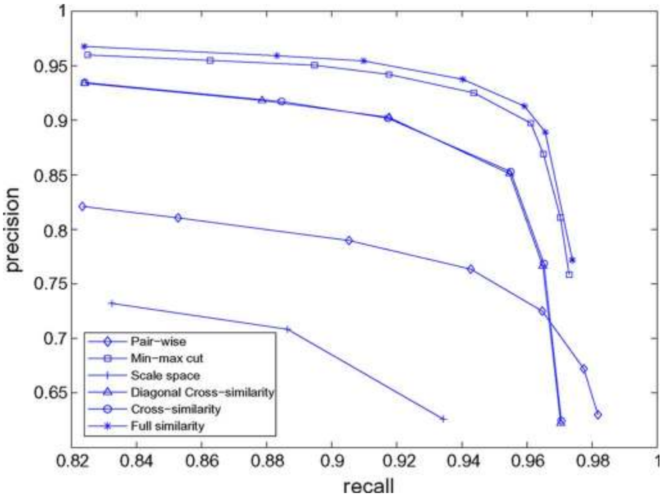


Fig. 11. Performance evaluation of the construction methods of continuity signal.

of Fig. 9. For GTs detection, four kinds of features are evaluated. The experimental setting is the same to the "mrs_late_{135}" as explained in the following Section V-B-2. By adjusting the ratio of misclassification penalties between positive and negative examples, the *recall* versus *precision* curves can be obtained as shown in (b) of Fig. 9. The result of CUTs detection indicates that the color space and quantization schemes affect the performance little, while the block based features outperform the global ones. That is because the global feature is not distinctive enough to capture the transitions between the same scene. When the number of blocks increases from 16 to 64, the performance does not show obvious improvement. The result of GTs detection shows that when the feature varies from coarse to fine, the performance first increases then decreases. The **RGB4_48** achieves the best performance. We investigate the outputs of various features and find that with finer features such as **RGB16_48** and **RGB64_48** the system is too sensitive to the disturbances of motion. That is why the performance of **RGB64_48** is worse than that of **RGB4_48**. This experiment has shown the importance of the tradeoff between invariance and sensitivity. In the following experiments, the "**RGB4_48**" feature is adopted.

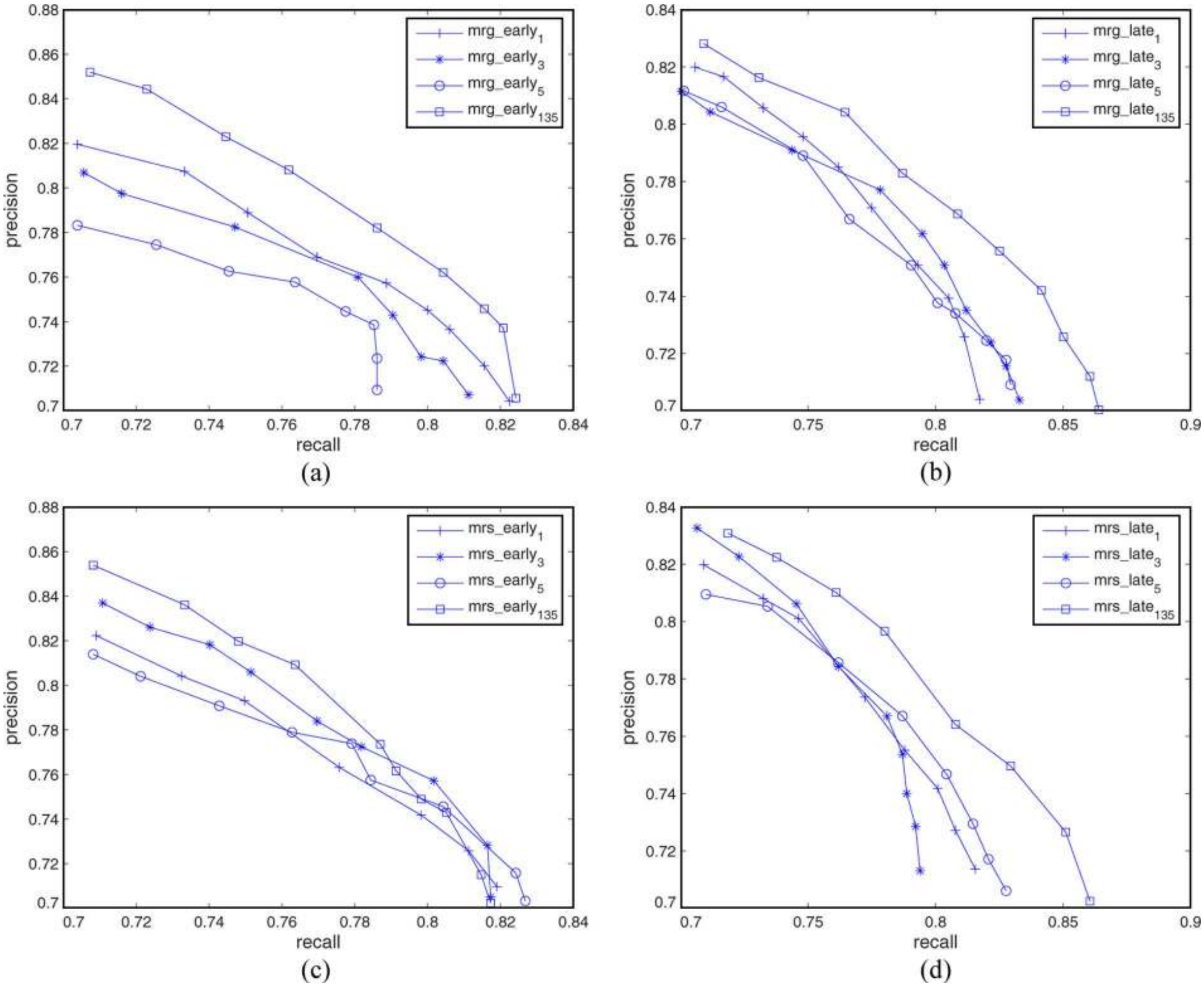
2) Continuity Signal Construction: Besides graph partition model, five other related approaches are implemented to perform CUTs detection and evaluated on both the "**D2003_NO_GT**" and "**D2004_NO_GT**" collections. They are the following.

Pair-wise	Pair-wise comparison of the successive frames.
Min-max cut	The algorithm proposed in Section IV.
Scale space	Kernel correlation [13] by Fig. 10(a).
Diagonal CS	Kernel correlation [13] by Fig. 10(b).
Cross S	Kernel correlation [13] by Fig. 10(c).
Full S	Kernel correlation [13] by Fig. 10(d).

Each of the above methods yields a curve of continuity signal. For each curve, a global threshold is employed to determine whether a CUT occurs. Varying the thresholds, the corresponding *precision* versus *recall* curves are obtained, as shown in the Fig. 11. The "**Pair-wise**" performs worst. It does not incorporate the contextual information while constructing the continuity signal. It can not successfully distinguish CUTs from other disturbances like flashlights. All the other five approaches via multipair comparison outperform the pair-wise comparison method. The score of "**Scale space**" kernel correlation actually is the mean of the frame differences in the neighborhood of the current position. With this smoothing effect, a lot of local minima on the curve are eliminated, and thus the "**Scale space**" gets a better performance compared to "**Pair-wise**" method. Both the "**Diagonal CS**" and the "**Cross S**" kernels emphasize the dissimilarity between the different shots. They evaluate the current score by incorporating multipair comparisons of the frames before and after the position. They performs almost the same, but the "**Diagonal CS**" is more efficient since fewer comparisons are performed. The "**Full S**" somewhat outperforms the proposed "**Min-max cut**" method and both of them perform best. This is not surprising, since they both consider the similarity between different shots and within the the same shot. In the perspective of graph partition model, the "**Full S**" is an alternative definition of min-max cut objective

$$M\mathrm{cut}(A,B) = \mathrm{assoc}(A) + \mathrm{assoc}(B) - 2 \times \mathrm{cut}(A,B). \quad (23)$$

Note that our experimental result is inconsistent with that of Cooper [13], in which the author claims that the "**Full S**" performs worst. However, we believe that our results are more reliable. First, we evaluate them on a more delicate data set without GTs. Second, we design a more straightforward experimental setup, in which a simple histogram feature and thresholding method are adopted, while Cooper has employed multiscale features and KNN to classify CUTs and non-CUTs.



(a) (b) (c) (d)

Fig. 12. Evaluation of multiresolution analysis and fusion methods in four different experimental settings: (a) "mrg" construction of the feature vectors and early fusion strategy. (b) "mrg" construction of the feature vectors and late fusion strategy. (c) "mrs" construction of the feature vectors and early fusion strategy. (d) "mrs" construction of the feature vectors and late fusion strategy. For each setting, four approaches are compared, and the recall versus precision curve of the GT detection of each approach is depicted.

3) Multiresolution Analysis: Here the experiments aim to figure out whether the multiresolution analysis is effective while detecting GTs, and further which setting of construction methods and fusion strategies is the most effective one. LibSVM is adopted to train the C-SVC model of the GTs detection [61]. Radial basis function (RBF) kernel is used in the SVMs model, and the best parameter settings are chosen via cross-validation processes. The models are trained on the "**D2003**" and the "**D2004**" collections, and then are tested on the "**D2005**" collections. The two construction methods of multiresolution feature vectors, i.e., **M**ulti-**R**esolution **G**raph (abbreviated as "mrg") and **M**ulti-**R**esolution **S**core (abbreviated as "mrs") are implemented. For each construction method, the two fusion strategies, namely early fusion and late fusion, are evaluated. The name of each experimental setting is represented as the combination of construction methods, fusion methods and resolutions adopted. For example, with "mrs_late_1," we mean that the "mrs" construction method, late fusion and $\delta = 1$ are adopted, but with "mrg_early_{135}," we mean the "mrg" construction method, early fusion and $\delta \in \{1, 3, 5\}$ are adopted.

While training the SVMs model, by adjusting the ratio of misclassification penalties between positive and negative examples, the *recall* versus *precision* of the output can be controlled. Concretely, with lower penalty ratio, the *precision* is preferred to *recall*, otherwise, with higher penalty ratio, the *recall* is preferred to *precision*. The comparison performances of the four experimental settings are depicted in Figs. 12 and 13. As shown in the Fig. 12, in each setting, the fusion of $\delta \in \{1, 3, 5\}$ resolutions outperforms the methods with single resolution, which shows that the multiresolution analysis and fusion is effective to boost the performance of GT detection. As the Fig. 13 shows, the four settings of multiresolution analysis yield almost the equal performances. On the one hand, with the same construction method, the late fusion method will yield *recall*-leaning result compared to the early fusion one. On the other hand, with the same fusion method, the two construction methods yield almost the same areas under the *recall* versus *precision* curves. Therefore, the construction methods do not cause distinct differences. The fusion strategy can be used to handle the tradeoff between *recall* and *precision*. If high *recall* is preferred, the late fusion can be adopted, otherwise, the early fusion is preferred.

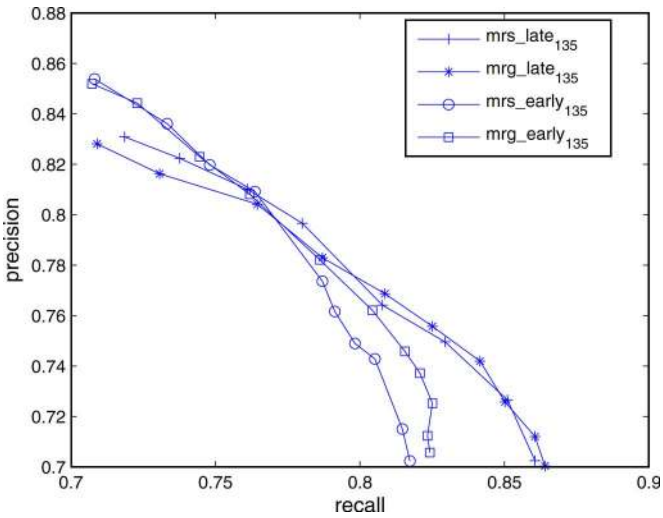


Fig. 13. Evaluation of the four different experimental settings. For each setting, the $\delta \in \{1, 3, 5\}$ resolutions are adopted, and the recall versus precision curve of the GT detection of each approach is depicted.

4) SVMs Active Learning: The objective of the experiments here is two-fold. First, the comparison between thresholding and SVMs is carried out to find out whether the machine learning method outperforms the ad hoc thresholding scheme. Second, the comparison between nonactive learning and active learning is performed to evaluate the effectiveness of active learning. Therefore, three approaches are implemented and compared:

Threshold	Global threshold for classification.
SVM_random	$T_{\mathrm{CUT}} = 1.00$ in (15).
SVM_active	$T_{\mathrm{CUT}} = 0.94$ in (15).

The continuity signal is constructed by graph partition model, and the above three approaches are used to classify CUTs and non-CUTs, respectively. The SVMs models are trained on "**D2003_NO_GT**" and "**D2004_NO_GT**" collections. They are tested on the "**D2005_NO_GT**" collections. For the **Threshold** method, a global threshold is tuned to obtain the *recall* versus *precision* curve, and for the two SVMs-based methods, similar to the GT detection, the penalty ratio is tuned to get the *recall* versus *precision* curve. As shown in Fig. 14, both the SVMs-based approaches outperform the **Threshold** method. And the **SVM_active** method achieve almost equal performance to the **SVM_random** method. Furthermore, as shown in the Table II, the **SVM_active** requires fewer training examples than **SVM_random**. Therefore, on the one hand, the training efficiency of the active learning method outperforms the nonactive learning one. On the other hand, there are fewer support vectors in the model for active learning, which will speed up the classification efficiency.

5) System Evaluation: In 2005, the proposed system of Section IV participated in the evaluation of SBD of TRECVID 2005. Trained on the "**D2003**" and the "**D2004**" collections, and tested on the "**D2005**" collections, via adjusting the penalty ratio of SVMs, ten runs were submitted. Besides our system, 20 other systems all over the world participated in the evaluation. There were totally 165 runs submitted. The details of the other approaches and the evaluation results can be found in the online proceeding of the Workshop of TRECVID 2005 [6]. The top 30 results of the evaluation are summarized in Table III. Ranked by F_1 measure, all the ten runs whose "SysID" are with the prefix "thu" are in the top 30 runs in detecting all the transitions. According to the other evaluation criteria, our system is also among the best.

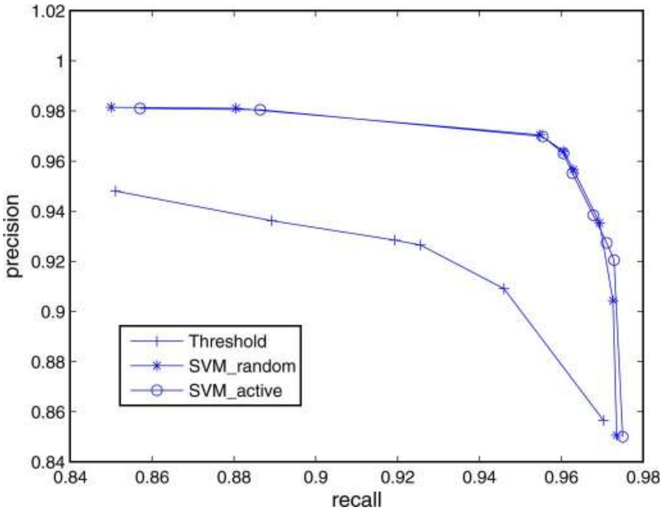


Fig. 14. Performance evaluation of SVMs active learning.

TABLE II
COMPARISON OF THE EFFICIENCIES BETWEEN NONACTIVE LEARNING AND ACTIVE LEARNING WITH SVMS

Name	# of Training Examples		Training	# of Support Vectors	
	Positive	Negative	Time	Positive	Negative
SVM_random	7352	57154	176.6s	457	7245
SVM_active	7305	6022	16.6s	723	780

VI. CONCLUSIONS AND FURTHER DISCUSSIONS

We have conducted a formal study of SBD problem in this paper. A general formal framework is proposed. Several major challenges to the framework are also identified. Furthermore, according to the formal framework, a comprehensive review of existing techniques is presented. The representative approaches are categorized and compared according to their roles in the formal framework. Optimal criteria for each module of the framework are also discussed, which will probably provide practical guide for developing novel methods. As an example, we present a unified SBD system based on graph partition model. Finally, we carry out extensive experiments on the platform of TRECVID. The experiments not only verify the optimal criteria identified above, but also show that the proposed system is among the best in the evaluation of TRECVID 2005.

In the above, SBD has been formulated in the pattern recognition perspective. The connection between SBD and some other pattern classification problems has been naturally established. Thus, they can benefit from each other. Here, we will present a rough discussion on what SBD can learn from similar problems of the related fields. The three mappings identified by the

TABLE III
SBD RESULTS AT TRECVID 2005. TOP 30 (OUT OF 165) RUNS AT EACH OF THE FOUR EVALUATION MEASURES ARE LISTED IN DECREASING F_1 MEASURE ORDER. THE "SysID" OF THE TEN RUNS OF OUR SYSTEM IS WITH THE PREFIX "thu," AND THE "SysID" OF THE OTHER RUNS ARE MARKED WITH NIST MARKER CONVENTIONS. THE RUNS WITH THE ID OF "thu" ARE HIGHLIGHTED WITH GRAY CELL BACKGROUND

	All Transitions				Cuts				Gradual Transitions				Gradual Frame Accuracy			
	SysID	Rcl	Prc	F#	SysID	Rcl	Prc	F#	SysID	Rcl	Prc	F#	SysID	Rcl	Prc	F#
1	thu26	0.894	0.901	0.897	bs-8	0.936	0.949	0.942	thu26	0.788	0.791	0.789	N208	0.833	0.871	0.852
2	thu12	0.904	0.890	0.897	bs-7	0.93	0.955	0.942	thu01	0.818	0.757	0.786	N207	0.827	0.877	0.851
3	thu02	0.912	0.878	0.895	bs-10	0.926	0.952	0.934	thu05	0.806	0.767	0.786	N212	0.827	0.875	0.850
4	thu01	0.901	0.887	0.894	bs-9	0.920	0.956	0.938	thu02	0.827	0.746	0.784	N202	0.824	0.878	0.850
5	thu25	0.872	0.916	0.893	thu25	0.930	0.941	0.935	thu12	0.771	0.789	0.784	thu13	0.880	0.821	0.849
6	thu13	0.882	0.903	0.892	thu13	0.949	0.921	0.935	thu07	0.838	0.733	0.782	N209	0.824	0.872	0.847
7	A15	0.903	0.881	0.892	thu26	0.930	0.939	0.934	A15	0.773	0.781	0.777	PS-8	0.830	0.864	0.847
8	A13	0.904	0.877	0.890	thu02	0.941	0.928	0.934	N204	0.838	0.722	0.776	thu25	0.876	0.818	0.846
9	A132s	0.902	0.878	0.890	thu12	0.949	0.919	0.934	thu09	0.854	0.710	0.775	PS-7	0.840	0.852	0.846
10	A16	0.899	0.880	0.889	thu01	0.929	0.936	0.932	N212	0.848	0.714	0.775	n194	0.823	0.870	0.846
11	A132	0.901	0.878	0.889	thu09	0.949	0.914	0.931	thu23	0.837	0.718	0.773	PS-10	0.798	0.898	0.845
12	thu09	0.925	0.856	0.889	A15	0.947	0.914	0.930	PS-6	0.688	0.881	0.773	N198	0.812	0.879	0.844
13	thu07	0.888	0.886	0.887	A13	0.947	0.914	0.930	n207	0.842	0.712	0.772	N197	0.812	0.879	0.844
14	thu05	0.920	0.854	0.886	A132s	0.946	0.914	0.930	A11	0.774	0.769	0.771	PS-6	0.849	0.837	0.84
15	A06	0.881	0.888	0.884	A132	0.946	0.914	0.930	PS-5	0.705	0.850	0.771	N210	0.818	0.869	0.843
16	A11	0.888	0.881	0.884	A16	0.946	0.913	0.929	A13	0.776	0.765	0.770	PS-9	0.805	0.883	0.842
17	A09	0.889	0.880	0.884	A06	0.941	0.915	0.928	A132s	0.771	0.769	0.770	thu12	0.859	0.824	0.841
18	thu23	0.927	0.845	0.884	thu07	0.905	0.948	0.926	A132	0.771	0.769	0.770	PS-5	0.861	0.822	0.841
19	A04s	0.879	0.878	0.879	A09	0.928	0.921	0.924	a16	0.76	0.78	0.770	thu26	0.860	0.821	0.840
20	A04	0.878	0.877	0.878	A11	0.928	0.920	0.924	A09	0.773	0.761	0.767	thu05	0.847	0.831	0.839
21	N204	0.909	0.845	0.876	thu23	0.957	0.892	0.923	N209	0.843	0.700	0.765	N204	0.836	0.837	0.837
22	N212	0.914	0.839	0.875	it1	0.917	0.929	0.923	sys10	0.741	0.790	0.765	thu02	0.845	0.828	0.836
23	N207	0.912	0.840	0.875	A04s	0.943	0.901	0.922	N208	0.848	0.695	0.764	thu07	0.846	0.825	0.835
24	PS-4	0.863	0.880	0.871	A04	0.943	0.901	0.921	thu25	0.701	0.831	0.760	thu01	0.841	0.829	0.835
25	N208	0.913	0.832	0.871	thu05	0.959	0.883	0.919	sys07	0.765	0.756	0.760	A15	0.772	0.909	0.835
26	N209	0.916	0.826	0.867	bs-1	0.932	0.896	0.914	sys06	0.756	0.764	0.760	A04s	0.766	0.915	0.834
27	PS-5	0.826	0.913	0.867	bs-3	0.929	0.898	0.913	N202	0.848	0.688	0.760	A04	0.766	0.915	0.834
28	bs-10	0.871	0.86	0.865	N207	0.936	0.890	0.912	N210	0.853	0.683	0.759	PS-4	0.871	0.799	0.833
29	bs-9	0.867	0.862	0.864	bs-5	0.914	0.910	0.912	sys06	0.769	0.748	0.758	A06	0.774	0.901	0.832
30	N210	0.915	0.817	0.863	g8	0.924	0.900	0.912	N198	0.850	0.684	0.758	A13	0.769	0.907	0.832

formal framework are in fact the core research problems of pattern recognition, which have undergone relatively mature evolution. First, for example, the methods of visual content representation and similarity measure have been thoroughly investigated in the field of content-based image retrieval (CBIR) [62], yet only few of them have been tried and evaluated in the problem of SBD. Second, via the construction of continuity signal, video sequence is transformed from a three dimension signal to a one dimension signal. The shot transitions are identified by the recognition of the shape of the one dimension signal. Similar problems exist in the related fields, such as temporal data segmentation [63], signal segmentation [64], and image segmentation [65]. Take image segmentation for example, it has attracted intensive research in the field of computer vision. Various approaches, such as JSEG [66], Mean Shift [67], and graph partition model [53], [54], [68], have been proposed. The principles underlying these techniques can be transformed to serve the purpose of SBD. In fact, the method introduced in Section IV is originally inspired by the graph-based image segmentation method of normalized cut [54]. Finally, the statistical machine learning approaches have been popular and shown some superiority in other pattern recognition problems. However, in the field of SBD, the efforts to replace thresholding by machine learning have begun only recently. More importantly, machine learning perhaps will provide powerful tools of information fusion for multimodalities SBD techniques. The importation of these ideas may be novel drives to the advance of SBD.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful suggestions. They also thank Dr. C.-J. Lin for the code of *libSVM* and National Institute of Standards and Technology (NIST) for providing TRECVID data sets and the evaluation tools, without which this work might not have been possible.

REFERENCES

[1] N. Dimitrova, H. J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video content analysis and retrieval," *IEEE Multimedia*, vol. 9, no. 3, pp. 42–55, Sep. 2002.

[2] L. A. Rowe and R. Jain, "Acm sigmm retreat report on future directions in multimedia research," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 1, no. 1, pp. 3–13, Feb. 2005.

[3] S. W. Smoliar and H.-J. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia*, vol. 1, no. 2, pp. 62–72, Jun. 1994.

[4] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video abstracting," *Commun. ACM*, vol. 40, no. 12, pp. 55–62, Dec. 1997.

[5] V. Kobla, D. DeMenthon, and D. Doermann, "Special effect edit detection using videotrails: a comparison with existing techniques," in *Proc. SPIE Conf. Storage Retrieval Image Video Databases VII*, Jan. 1999, pp. 302–313.

[6] *NIST, Homepage of Trecvid Evaluation.* [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[7] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 3–19, Jan. 2000.

[8] R. Lienhart, "Reliable transition detection in videos: a survey and practitioner's guide," *Int. J. Image Graph.*, vol. 1, no. 3, pp. 469–486, 2001.

[9] A. Hanjalic, "Shot boundary detection: unraveled and resolved?," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 90–105, Feb. 2002.

[10] M. Albanese, A. Chianese, V. Moscato, and L. Sansone, "A formal model for video shot segmentation and its application via animate vision," *Multimedia Tools Appl.*, vol. 24, no. 3, pp. 253–272, 2004.

[11] J. Bescós, G. Cisneros, J. M. Martínez, J. M. Menendez, and J. Cabrera, "A unified model for techniques on video shot transition detection," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 293–307, Apr. 2005.

[12] J. Yuan, J. Li, F. Lin, and B. Zhang, "A unified shot boundary detection framework based on graph partition model," in *Proc. ACM Multimedia 2005*, Nov. 2005, pp. 539–542.

[13] M. Cooper, "Video segmentation combining similarity analysis and classification," in *Proc. ACM Multimedia 2004*, Oct. 2004, pp. 252–255.

[14] Y. Qi, A. Hauptmann, and T. Liu, "Supervised classification for video shot segmentation," in *IEEE Conf. Multimedia Expo*, Jul. 2003, vol. 2, pp. 689–692.

[15] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, pp. 1–13, Feb. 2000.

[16] T. Volkmer, S. M. M. Tahaghoghi, and H. Williams, "RMIT university at trecvid 2004," in *Proc. TRECVID 2004 Workshop*, 2004 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/rmit.ps

[17] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *Proc. ACM Multimedia 1996*, Nov. 1996, pp. 65–73.

[18] B. Janvier, E. Bruno, S. Marchand-Maillet, and T. Pun, "Information-theoretic framework for the joint temporal partioning and representation of video data," in *European Conf. Content-Based Multimedia Indexing (CBMI03)*, 2003.

[19] T. Mitchell, *Machine Learning*. New York: McGraw Hill, 2005, ch. 1 [Online]. Available: http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

[20] R. Lienhart, "Reliable dissolve detection," in *Proc. SPIE Storage Retrieval Media Database*, Jan. 2001, vol. 4315, pp. 219–230.

[21] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," in *Proc. SPIE Storage Retrieval Image Video Databases IV*, Jan. 1996, vol. 2664, pp. 170–179.

[22] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Proc. SPIE Image Video Process. VII*, Jan. 1999, vol. 3656, pp. 290–301.

[23] S. Lefèvre, J. Holler, and N. Vincent, "A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval," *Real-Time Imag.*, vol. 9, no. 1, pp. 73–98, Feb. 2003.

[24] T. Kikukawa and S. Kawafuchi, "Development of an automatic summary editing system for the audio visual resources," *Trans. IEICE*, vol. J75-A, no. 2, pp. 204–212, 1992.

[25] S. K. Choubey and V. V. Raghavan, "Generic and fully automatic content-based image retrieval using color," *Pattern Recog. Lett.*, vol. 18, no. 11–13, pp. 1233–1240, 1997.

[26] H.-J. Zhang, C. Y. Low, and S. W. Smoliar, "Video parsing and browsing using compressed data," *Multimedia Tools Appl.*, vol. 1, no. 1, pp. 89–111, 1995.

[27] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks," in *Proc. ACM Multimedia*, San Francisco, CA, Nov. 1995, pp. 189–200.

[28] P. Bouthemy, M. Gelgon, and F. Ganansia, "A unified approach to shot change detection and camera motion characterization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 7, pp. 1030–1044, Oct. 1999.

[29] M. G. Chung, H. Kim, and S. M.-H. Song, "A scene boundary detection method," in *Proc. Int. Conf. Image Processing*, Sep. 2000, vol. 3, pp. 933–936.

[30] S. J. F. Guimar and M. Couprie, "Video segmentation based on 2d image analysis," *Pattern Recognit. Lett.*, vol. 24, no. 7, pp. 947–957, 2003.

[31] C.-W. Ngo, T.-C. Pong, and R. T. Chin, "Video partitioning by temporal slice coherency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 8, pp. 941–953, Aug. 2001.

[32] M. Ahmed, A. Karmouch, and S. Abu-Hakima, "Key frame extraction and indexing for multimedia databases," in *Proc. Vis. Interface Conf.*, 1999, pp. 506–511.

[33] S. H. Kim and R.-H. Park, "Robust video indexing for video sequences with complex brightness variations," in *Proc. IASTED Int. Conf. Signal Image Process.*, Kauai, HI, Aug. 2002, pp. 410–414.

[34] W. J. Heng and K. N. Ngan, "High accuracy flashlight scene determination for shot boundary detection," *Signal Process.: Image Commun.*, vol. 18, no. 3, pp. 203–219, Mar. 2003.

[35] M. Leszczuk and Z. Papir, "Accuracy versus speed tradeoff in detecting of shots in video content for abstracting digital video libraries," in *Lecture Notes In Computer Science*. London, U.K.: Springer-Verlag, 2002, vol. 2515, pp. 176–189.

[36] W. Zheng, J. Yuan, H. Wang, F. Lin, and B. Zhang, "A novel shot boundary detection framework," in *Proc. SPIE Vis. Commun. Image Process.*, Jun. 2005, vol. 5960, pp. 410–420.

[37] S.-C. Jun and S.-H. Park, "An automatic cut detection algorithm using median filter and neural network," in *Proc. ITC-CSCC2000*, Jul. 2000, pp. 1049–1052.

[38] M. Slaney, D. Ponceleon, and J. Kaufman, "Multimedia edges: finding hierarchy in all dimensions," in *Proc. ACM Multimedia*, Sep. 2001, pp. 29–40.

[39] M. J. Pickering, D. Heesch, R. O'Callaghan, S. Rger, and D. Bull, "Video retrieval using global features in keyframes," in *Proc. TREC Video Track*, 2002 [Online]. Available: http://trec.nist.gov//pubs/trec11/papers/imperial.pickering.pdf

[40] T. Truong, C. Dorai, and S. Venkatesh, "New enhancements to cut, fade, and dissolve detection processes in video segmentation," in *Proc. ACM Multimedia*, Los Angeles, CA, 2000, pp. 219–227.

[41] B.-L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 533–544, Dec. 1995.

[42] M. R. Naphade, R. Mehrotra, A. Ferman, J. Warnick, T. S. Huang, and A. M. Tekalp, "A high-performance shot boundary detection algorithm using multiple cues," in *IEEE Inte. Conf. Image Process.*, 1998, pp. 884–887.

[43] C.-W. Ngo, "A robust dissolve detector by support vector machine," in *Proc. ACM Multimedia*, 2003, pp. 283–286.

[44] T.-S. Chua, H. Feng, and C. A. , "An unified framework for shot boundary detection via active learning," in *Proc. ICASSP*, Hong Kong, Apr. 2003, vol. 2, pp. 845–848.

[45] H. Feng, W. Fang, S. Liu, and Y. Fang, "A new general framework for shot boundary detection and key-frame extraction," in *Proc. 7th ACM SIGMM Int. Workshop Multimedia Inf. Retrieval*, 2005, pp. 121–126.

[46] U. Naci and A. Hanjalic, "TU Delft at TRECVID 2005: Shot boundary detection," in *Proc. TRECVID 2005 Workshop*, 2005 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tuDelft.pdf

[47] A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proc. ACM Multimedia*, 1994, pp. 357–364.

[48] A. M. Alattar, "Detecting and compressing dissolve regions in video sequences with a DVI multimedia image compression algorithm," in *Proc. IEEE ISCAS*, May 1993, vol. 1, pp. 13–16.

[49] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst.*, vol. 1, no. 1, pp. 10–28, Jun. 1993.

[50] Y. Lin, M. S. Kankanhalli, and T.-S. Chua, "Temporal multiresolution analysis for video segmentation," in *Proc. SPIE Conf. Storage Retrieval Media Database VIII*, 2000, vol. 3972, pp. 494–505.

[51] J. Yuan, B. Zhang, and F. Lin, "Graph partition model for robust temporal data segmentation," in *Proc. Adv. Knowledge Discov. Data Mining: 9th Pacific-Asia Conf., PAKDD 2005*, May 2005, pp. 758–763.

[52] J. Yuan, L. Xiao, D. Wang, D. Ding, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin, and B. Zhang, "Tsinghua University at TRECVID 2005," in *Proc. TRECVID Workshop 2005*, Nov. 2005 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tsinghua.pdf

[53] S. Wang and J. Mark, "Image segmentation with ratio cut," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 6, pp. 675–690, Jun. 2003.

[54] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[55] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *Proc. 2001 IEEE Int. Conf. Data Mining*, 2001, pp. 107–114.

[56] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[57] G. Schohn and D. Cohn, "Less is more: active learning with support vector machines," in *Proc. 17th Int. Conf. Mach. Learning*, 2000, pp. 839–846.

[58] G. Iyengar, H. J. Nock, and C. Neti, "Discriminative model fusion for semantic concept detection and annotation in video," in *Proc. 11th ACM Int. Conf. Multimedia*, Nov. 2003, pp. 255–258.

[59] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, Nov. 2005, pp. 399–402.

[60] L. Zhang, F. Lin, and B. Zhang, "A CBIR method based on color-spatial feature," in *IEEE Reg. 10 Annu. Int. Conf. (TENCON'99)*, Cheju, Korea, 1999, vol. 1, pp. 166–169.

[61] C.-J. L. , C.-W. Hsu, and C.-C. Chang, A Practical Guide to Support Vector Classification 2005 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/, Tech. Rep.

[62] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[63] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: a survey and novel approach," in *Data Mining in Time Series Databases*. Singapore: World Scientific, 2003.

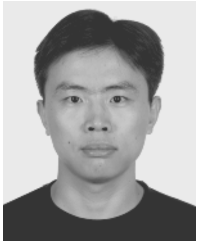
[64] M. Basseville and I. V. Nikiforov*, Detection of Abrupt Changes: Theory and Application*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[65] J. F. Canny, "Finding edges and lines in images," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1983.

[66] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[67] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[68] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

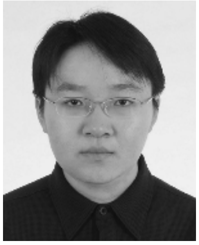


Jinhui Yuan received the B.E. degree in computer science and technology from Xidian University, Xidian, China, in 2003. He is currently working toward the Ph.D degree in the Department of Computer Science and Technology, Tsinghua University, Beijing, China.

He works with the Intelligent Multimedia Group, State Key Laboratory of Intelligent Technology and Systems, Beijing, China. His major research interests include content based video retrieval, pattern recognition, and machine learning.

Huiyi Wang received the B.E. degree in computer science and technology from Tsinghua University, Beijing, China, in July, 2006, where she is currently working toward the M.S. degree in the Department of Computer Science and Technology.

She has been participating in research on shot boundary detection and automatic video search system at the Intelligent Multimedia Group since 2004. Her major research interests include content-based video retrieval and machine learning.



Lan Xiao received the B.E. degree in computer science from Tsinghua University, Beijing, China, in 2006. She is currently working toward the graduate degree at the Computer Science Department, Columbia University, New York.

She has been participating in the research on video shot boundary detection and interactive video retrieval system at the Intelligent Multimedia Group, State Key Laboratry of Intelligent Technology and Systems since 2005.

Wujie Zheng received the B.E. degree in computer science and technology from Tsinghua University, Beijing, China, in 2004, where he is currently working toward the M.S. degree in the Department of Computer Science and Technology.

He works with the Intelligent Multimedia Group, State Key Laboratory of Intelligent Technology and Systems, Beijing, China. His major research interests include content based video retrieval and shot boundary detection.

Jianmin Li received the B.E/ degree in computer science and technology and the Ph.D. degree in computer application from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 1995 and 2003, respectively.

Currently, he is an assistant Researcher of Department of Computer Science and Technology, Tsinghua University. His main research interests include machine learning, speech synthesis, and information retrieval. He has published more than ten papers.

Fuzong Lin graduated from the Department of Automatic Control, Tsinghua University, Beijing, China, in 1970.

He is a Professor in the Department of Computer Science and Technology, Tsinghua University. His research interests include multimedia information processing, web-based learning, and teaching technologies.

Bo Zhang graduated from the Department of Automatic Control, Tsinghua University, Beijing, China, in 1958.

Currently, he is a Professor in the Department of Computer Science and Technology, Tsinghua University and a Fellow of Chinese Academy of Sciences, Beijing, China. His main interests are artificial intelligence, pattern recognition, neural networks, and intelligent control. He has published over 150 papers and four monographs in these fields.