Chapter IV

Statistical Audio-Visual Data Fusion for Video Scene

Vyacheslav Parshin, Ecole Centrale de Lyon, France
Liming Chen, Ecole Centrale de Lyon, France

Abstract

Automatic video segmentation into semantic units is important in order to organize an effective content-based access to long video. In this work, we focus on the problem of video segmentation into narrative units called scenes—aggregates of shots unified by a common dramatic event or locale. In this work, we derive a statistical video scene segmentation approach that detects scenes boundaries in one pass, fusing multi-modal audiovisual features in a symmetrical and scalable manner. The approach deals properly with the variability of real-valued features and models their conditional dependence on the context. It also integrates prior information concerning the duration of scenes. Two kinds of features extracted in visual and audio domain are proposed. The results of experimental evaluations carried out on ground truth video are reported. They show that our approach effectively fuses multiple modalities with higher performance compared with an alternative rule-based fusion technique.

Introduction

A constantly growing amount of available digitized video stored at centralized libraries or even on personal computers gives rise to the need for an effective means of navigation that allows a user to locate a video segment of interest. Searching of such a segment sequentially using simple fast-forward or fast-reverse operations provided by most of the existing players is tedious and time-consuming. A content-based access could greatly simplify this task, giving to a user the possibility to browse a video organized as a sequence of semantic units. Such an organization also could facilitate the task of automatic video retrieval, restricting the search by the scope of meaningful semantic segments. Another potential area of application is an automatic generation of video summaries or skims that preserve the semantic organization of the original video.

As the basic building blocks of professional video are shots—sequences of contiguous frames recorded from a single camera—it is natural to divide a video into these units. Unfortunately, the semantic meaning they provide is at too low of a level. Common video of about one or two hours (e.g., a full-length film) usually contains hundreds or thousands of shots—too many to allow for efficient browsing. Moreover, individual shots rarely have complete narrative meaning. Users are more likely to recall whole dramatic events or episodes, which usually consist of several contiguous shots. In this work, we consider the task of automatic segmentation of narrative films, such as most movies, into something more meaningful than shots—high-level narrative units called scenes, or aggregates of shots unified by a common dramatic event or locale. We need shot segmentation at the first preliminary processing step since scenes are generated as groups of shots. Segmentation into scenes can be considered the next level of content generation, yielding a hierarchical semantic structure of video in which shots are preserved to form the lower level. In this work, we are not concerned with the problem of shot segmentation or adopting one of already existing techniques (Boresczky & Rowe, 1996; Lienhart, 1999) but rather focus on the task of video segmentation into scenes.

Sharing a common event or locale, shots of a scene usually are characterized by a similar environment that is perceivable in both the visual and audio domains. So, both the image sequence and the audio track of a given video can be used to distinguish scenes. Since the same scene of a film usually is shot in the same settings by the same cameras that are switched repeatedly, it can be detected from the image track as a group of visually similar shots. The visual similarity is established using low-level visual features such as color histograms or motion vectors (Kender & Yeo, 1998; Rasheed & Shah, 2003; Tavanapong & Zhou, 2004). On the other hand, a scene transition in movie video usually entails abrupt changes of some audio features caused by a switch to other sound sources and sometimes by film editing effects (Cao, Tavanapong, Kim, & Oh, 2003; Chen, Shyu, Liao, & Zhang, 2002; Sundaram & Chang, 2000). Hence, sound analysis provides useful information for scene segmentation as well. Moreover, additional or alternative features can be applied. For example, editing rhythm, which usually is preserved during a montage of a scene, can be used to distinguish scenes as groups of shots of predictable duration (Aigrain, Joly, & Longueville, 1997); classification of shots into exterior or interior ones would allow for their grouping into the appropriate scenes (Mahdi, Ardebilian, & Chen, 1998), and so forth.

In order to provide reliable segmentation, there is a need to properly combine these multiple modalities to compensate for their inaccuracies. The common approach uses a set of rules according to which one source of information usually is chosen as the main one to generate initial scene boundaries, while the others serve for their verification (Aigrain, Joly, & Longueville, 1997; Cao, Tavanapong, Kim, & Oh, 2003) or further decomposition into scenes (Mahdi, Ardebilian, & Chen, 2000). Rules-based techniques, however, are convenient for a small number of features, generally do not take into account fine interaction between them, and are hardly extensible. Another frequent drawback of the existing methods is binarization of real-valued features that often leads to losses of information.

In this work, we derive a statistical scene segmentation approach that allows us to fuse multiple information sources in a symmetrical and flexible manner and is easily extensible to new ones. Two features are developed and used as such sources: video coherence that reveals possible scene changes through comparison of visual similarity of shots and audio dissimilarity reflecting changes in the audio environment. For the moment, we fuse these two types of information, but our approach easily can be extended to include additional data. In contrast to the common rule-based segmentation techniques, our approach takes into account a various confidence level of scene boundary evidence provided by each feature.

In our earlier work (Parshin, Paradzinets, & Chen, 2005) we already proposed a simpler approach (referenced hereafter as maximum likelihood ratio method) for the same scene segmentation task. The advantage of the approach proposed in this work (referenced hereafter as sequential segmentation method) is that it is based on less restrictive assumptions about observable feature vectors, allowing for their conditional dependence from the context, and takes into consideration a nonuniform statistical distribution of scene duration.

We have evaluated the proposed technique using a database of ground-truth video, including four full-length films. The evaluation results showed a superior segmentation performance of our sequential segmentation technique with respect to the previous maximum likelihood ratio, one that in its turn outperforms a conventional rule-based, multi-modal algorithm.

The remainder of this chapter has the following organization. First, the background and related work section briefly describes prior work on the scene segmentation problem and introduces the basic ideas that facilitate distinguishing video scenes in the visual and audio domain; coupling of multi-modal evidence about scenes is discussed as well. In the next section, we derive our sequential segmentation approach and make the underlying assumptions. Then, in the Feature Extraction section, we derive our video coherence measure used to distinguish scenes in the visual domain and provide details on our audio dissimilarity feature. In the Experiments section, we report the results of the experimental evaluation of the proposed scene segmentation approach using ground-truth video data. Final remarks then conclude this work.

Background and Related Work

Video Scene Segmentation Using Visual Keys

The common approach to video scene segmentation in the visual domain exploits the visual similarity between shots, which stems from specific editing rules applied during film montage (Bordell & Thompson, 1997). According to these rules, video scenes usually are shot by a small number of cameras that are switched repeatedly. The background and often the foreground objects shot by one camera are mostly static, and hence, the corresponding shots are visually similar to each other. In the classical graph-based approach (Yeung & Yeo, 1996), these shots are clustered into equivalence classes and are labeled accordingly. As a result, the shot sequence of a given video is transformed into a chain of labels that identifies the cameras. Within a scene, this sequence usually consists of the repetitive labels. When a transition to another scene occurs, the camera set changes. This moment is detected at a cut edge of a scene transition graph built for the video. For example, a transition from a scene shot by cameras *A* and *B* to a scene taken from cameras *C* and *D* could be represented by a chain *ABABCDCD* in which the scene boundary would be pronounced before the first *C*. An analogous approach was proposed by Rui, Huang, and Mehrotra (1999) in which shots first were clustered into groups that then were merged into scenes. Tavanapong and Zhou (2004) in their ShotWeave segmentation technique use additional rules to detect specific establishment and reestablishment shots that provide a wide view over the scene setting at the beginning and the end of a scene. They also suggest using only specific regions of video frames to determine more robustly the intershot similarity.

To overcome the difficulties resulting from a discrete nature of the segmentation techniques based on shot clustering, such as their rigidity and the need to choose a clustering threshold, continuous analogues have been proposed. Kender and Yeo (1998) reduce video scene segmentation to searching of maxima or minima on a curve describing the behavior of a continuous-valued parameter called video coherence. This parameter is calculated at each shot change moment as an integral measure of similarity between two adjacent groups of shots based on a short-memory model that takes into consideration the limitation and preferences of the human visual and memory systems. Rasheed and Shah (2003) propose to construct a weighted undirected shot similarity graph and detect scene boundaries by splitting this graph into subgraphs in order to maximize the intra-subgraph similarities and to minimize the inter-subgraph similarities.

In this work, we propose a continuous generalization of the discrete clustering-based technique, which is analogous to the approach of Kender and Yeo (1998) in the sense that it yields a continuous measure of video coherence. This measure then is used in our multimodal segmentation approach as a visual feature providing a flexible confidence level of the presence or absence of a scene boundary at each point under examination; the lower this measure is, the more possible is the presence of a scene boundary (see Figure 2). In contrast to the video coherence of Kender and Yeo (1998), which is a total sum of intershot similarities, our measure integrates only the similarity of the shot pairs that possibly are taken from the same camera.

Video Scene Segmentation in the Audio Domain

As the physical setting of a video scene usually remains fixed or changes gradually (when, for instance, the cameras follow moving personages), the sources of the ambient sound rest stable or change their properties smoothly and slowly. A scene change results in a shift of the locale, and hence, the majority of the sound sources also changes. This change can be detected as the moment of a drastic change of audio parameters characterizing the sound sources.

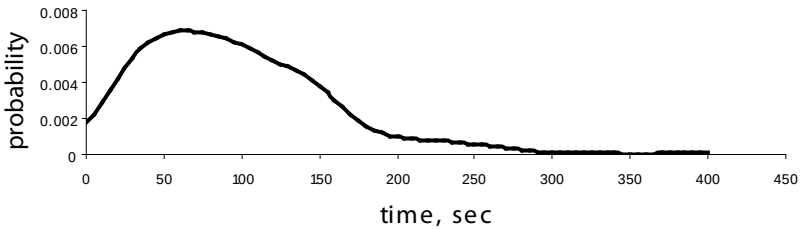
Since short-term acoustic parameters often are not capable of properly representing the sound environment (Chen, Shyu, Liao, & Zhang, 2002), these parameters often are combined within a long-term window. The resulting characteristics are evaluated within two adjacent time windows that adjoin a point of potential scene boundary (usually shot breaks) or its immediate vicinity (as sound change sometimes shifted by a couple of seconds during montage to create an effect of interscene connectivity) and then compared. A scene boundary is claimed if the difference is large enough. Sundaram and Chang (2000) model the behavior of various short-term acoustic parameters, such as cepstral flux, zero crossing rate, and so forth, with correlation functions that characterize the long-term properties of the sound environment. A scene change is detected when the decay rate of the correlation functions, the total for the all acoustic parameters, reaches a local maximum, as it means low correlation between these parameters caused by the change of the sound sources. Cao, Tavanapong, Kim, and Oh (2003) approximate long-term statistical properties of short-term acoustic parameters using normal distribution. At a potential scene boundary, these properties are compared by applying a weighted Kullback-Leibler divergence distance.

In this work, we adopt Kullback-Leibler distance as an audio dissimilarity feature providing the evidence of the presence or absence of a scene boundary in the audio domain. This distance represents the divergence between distributions of shot-term spectral parameters that are estimated using the continuous wavelet transform.

Multi-Modal Data Fusion

The common approach to segmentation of narrative video into scenes is based only on visual keys extracted from the image stream. In order to combine information extracted from the audio and image streams into one more reliable decision, a set of simple rules is usually applied. The audio stream can be used as an auxiliary data source to confirm or reject scene boundaries detected from the image sequence. For example, Cao, Tavanapong, Kim, and Oh (2003) first segment video into scenes in the visual domain and then apply sound analysis to remove a boundary of suspiciously short scenes, if it is not accompanied by a high value of audio dissimilarity. Jiang, Zhang, and Lin (2000) propose first to segment the video in the audio domain and to find potential scene boundaries at shot breaks accompanied by a change in the sound environment; these boundaries then are kept in the final decision if they are confirmed by low visual similarity between preceding and succeeding shots. Sundaram and Chang (2000) first segment video into scenes independently in the video and audio domains and then align visual and audio scene boundaries as follows. For visual and audio scene boundaries lying within a time ambiguity window, only the visual scene boundary is

Figure 1. Scene duration pdf



claimed to be the actual scene boundary; the rest of the boundaries are treated as the actual scene boundaries.

Rule-based approaches suffer from rigidity of the logic governing the feature fusion. Generally, each feature provides evidence about the presence or absence of a scene boundary with a different level of confidence, depending on its value. Making intermediate decisions, rule-based techniques ignore this difference for one or several features. Moreover, these techniques require the proper choice of thresholds, which usually are more numerous, the more rules that are applied. In this work, we derive a segmentation approach that fuses multiple evidences in a statistical manner, dealing properly with the variability of each feature. This approach is easily extensible to new features, in contrast to rule-based techniques that often become too complicated and cumbersome when many features are treated. We also take into consideration a nonuniform distribution of scene durations (see Figure 1) by including it as prior information.

Sequential Segmentation Approach

In this section, we derive our segmentation approach, which makes decisions about the presence or absence of a scene boundary at each candidate point. As scenes are considered as groups of shots, their boundaries occur at the moments of shot transitions. Therefore, in this work, these transitions are chosen as candidate points of scene boundaries. It is assumed that evidence about a scene boundary at an arbitrary candidate point is provided by a locally observable audiovisual feature vector. Further in this section, we first do some assumptions about observable features in the following subsection. Then an estimate of the posterior probability of scene boundary is derived, and the final segmentation algorithm is given in the next two subsections.

Conditional Dependence Assumptions about Observable Features

Let's consider an observable audiovisual feature vector D_i measured at a scene boundary candidate point i independently from the rest of vectors. In the general case, this vector is conditioned on the fact of presence or absence of a scene boundary not only at this point but at the neighboring points as well. Indeed, in the visual domain, the corresponding feature usually represents visual similarity between two groups of shots adjoining to the point under examination. If a scene boundary appears exactly between these groups, then the similarity measure usually has a local extremum. But if a scene boundary lies inside one of these groups, then the similarity measure takes an intermediate value that is closer to the extremum, the closer the scene boundary is (see Figure 2). Similar considerations also hold true for the audio data (see Figure 2).

For the purpose of simplification, we assume that local features are conditionally dependent on the distance to the closest scene boundary and are independent of the position of the rest of the scene boundaries. As the visual feature used in this work is a similarity measure applied to the whole shots, it is reasonable to assume the conditional dependence of this feature on the distance expressed in the number of shots. Let's denote a time-ordered sequence of scene boundaries as $B = \{b_1, b_2, ..., b_n\}$, in which each boundary is represented by the order number of the corresponding candidate point. As the scene boundary closest to an arbitrary candidate point i is one of two successive boundaries b_{k-1} and b_k surrounding this point so as $b_{k-1} \le i < b_k$, the likelihood of video feature v_i measured at point i given partitioning into scenes B can be written as:

$$P(v_i \mid B) = P(v_i \mid b_{k-1}, b_k) = P(v_i \mid \Delta_i), \tag{1}$$

in which Δ_i is the distance from point i to its closest scene boundary b_c defined as:

$$\Delta_i = i - b_c, \tag{2}$$

$$b_c = \begin{cases} b_{k-1}, \text{if } i - b_{k-1} \le b_k - i \\ b_k \text{ otherwise.} \end{cases} \tag{3}$$

The audio feature is defined in this work as a change in acoustic parameters measured within two contiguous windows of the fixed temporal duration. Therefore, we assume conditional dependence of this feature on the time distance to the closest scene boundary. Denoting the time of i-th candidate point as t_i, the temporal distance from point i to its closest scene boundary—as t_i, we write the likelihood of audio feature a_i measured at point i as:

$$P(a_i \mid B) = P(a_i \mid b_{k-1}, b_k) = P(a_i \mid \mathrm{t}_i), \tag{4}$$

in which

$$\tau_i = t_i - t_c, \tag{5}$$

$$t_c = \begin{cases} t_{b_{k-1}}, & t_i - t_{b_{k-1}} \le t_{b_k} - t_i \\ t_{b_k} \text{ otherwise.} \end{cases} \tag{6}$$

In this work, we calculate likelihood values $P(v_i \mid \Delta_i)$ and $P(a_i \mid \tau_i)$ using the corresponding probability density functions (pdf) considered to be stationary (i.e., independent of time index *i*). It is assumed that observable features are dependent on the closest scene boundary only if the distance to it is quite small (i.e., lower than some threshold that is on the order of the length of the time windows used to calculate these features). This assumption facilitates the learning of parameters of pdf estimates based on a set of learning data.

Taking into account expression (1) and (4), the likelihood of the total feature vector $D_i = \{v_i, a_i\}$ given partitioning into scenes *B* can be reduced to:

$$P(D_i \mid B) = P(D_i \mid b_{k-1}, b_k). \tag{7}$$

In this work, we assume conditional independence of the components of D_i given *B*:

$$P(D_i \mid B) = P(v_i \mid B)P(a_i \mid B) = P(v_i \mid b_{k-1}, b_k)P(a_i \mid b_{k-1}, b_k). \tag{8}$$

If more observable data are available, expression (8) can include additional feature vector components that provide an easy extensibility of our segmentation approach.

Segmentation Principles

Statistical analysis of scene duration shows that it has nonuniform distribution, as most scenes last from half a minute to two to three minutes (see Figure 1). In order to take into account the information about scene duration, we include a prior of a scene boundary that depends on the time elapsed from the previous scene boundary and does not depend on the earlier ones, much as it is done in the case of variable duration hidden Markov models (Rabiner, 1989). Furthermore, the posterior probability of a scene boundary b_k at point *i* is assumed to be conditionally dependent solely on local feature vector D_i given the position b_{k-1} of the previous scene boundary. This assumption agrees with the intuition that evidence of the presence or absence of a scene boundary at an arbitrary point is determined by the feature vector measured at the same point. Indeed, this feature vector reflects the degree of change in the visual and audio environment of a scene, and the larger this change is, the higher is

the probability of a scene change. Using Bayes rule, the posterior probability of *k*-th scene boundary at point *i* given b_{k-1} is written as:

$$P(b_k = i \mid D_i, b_{k-1}) = \frac{P(D_i \mid b_{k-1}, b_k = i)P(b_k = i \mid b_{k-1})}{P(D_i \mid b_{k-1}, b_k = i)P(b_k = i \mid b_{k-1}) + P(D_i \mid b_{k-1}, b_k \neq i)P(b_k \neq i \mid b_{k-1})}$$

$$= 1/\left[1 + \frac{P(D_i \mid b_{k-1}, b_k \neq i)P(b_k \neq i \mid b_{k-1})}{P(D_i \mid b_{k-1}, b_k = i)P(b_k = i \mid b_{k-1})}\right], \ \forall i > b_{k-1}. \quad (9)$$

In expression (9), it is further assumed that the next scene boundary b_{k+1} takes place a long time after boundary b_k, so that the likelihood of D_i given $b_k < i$ always is conditioned on b_k when computed according to expressions (1) through (6). We denote this assumption as $b_{k+1} = \infty$. It also is supposed that scene boundary duration is limited in time by a threshold value *S*. Then a possible position of *k*-th scene boundary is limited by a value m_k defined as:

$$m_k = \max\{l \mid t_l - t_{b_{k-1}} \leq S\}. \quad (10)$$

Under these assumptions, expression (9) is continued as:

$$P(b_k = i \mid D_i, b_{k-1}, b_{k+1} = \infty) =$$

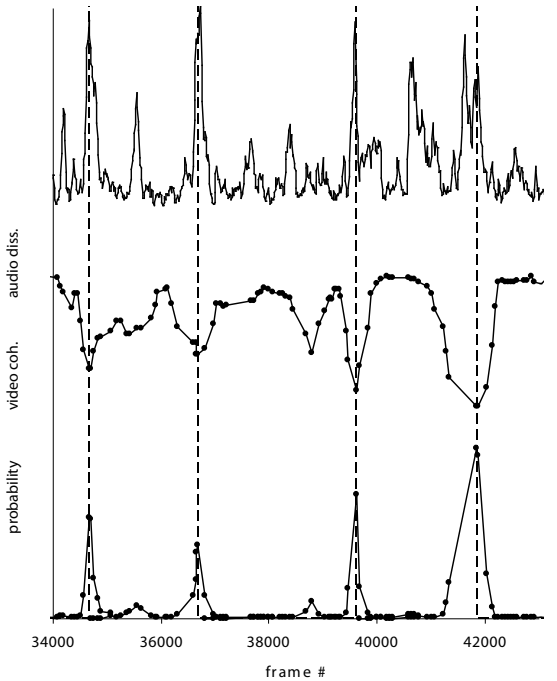
$$= 1/\left[1 + \frac{\sum_{l=b_{k-1}+1}^{i-1} P(D_i \mid b_k = l, b_{k+1} = \infty)P(b_k = l \mid b_{k-1}) + \sum_{l=i+1}^{m_k} P(D_i \mid b_{k-1}, b_k = l)P(b_k = l \mid b_{k-1})}{P(D_i \mid b_k = i)P(b_k = i \mid b_{k-1})}\right]. \quad (11)$$

It is assumed that the prior probability $P(b_k \mid b_{k-1})$ of scene boundary b_k is determined by the duration of the scene, which ends up at this boundary and is calculated using pdf of scene duration p_s as:

$$P(b_k \mid b_{k-1}) = \alpha p_s(t_{b_k} - t_{b_{k-1}}). \quad (12)$$

Normalizing coefficient α can be omitted when this expression is substituted in equality (11), as only the ratio of probability values is taken into account. In this work, we use a nonparametrical estimate of pdf p_s with Gaussian kernel (Duda & Hart, 1973) and limit its range of definition by lower and upper boundaries.

Figure 2. Audio dissimilarity (upper curve), video coherence (middle curve), and scene boundary posterior probability in sequential segmentation approach (partially overlapping curves in the bottom) vs. frame number; vertical dashed lines delimit scenes



We propose to segment an input video into scenes sequentially, choosing each next scene boundary based on the position of the previous one. So, the video can be segmented in real time with a time delay of the order of the maximal scene duration S. Knowing the position of scene boundary b_{k-1}, we select the next boundary b_k using the posterior probability estimated at each candidate point i, $i > b_{k-1}$, on time length S according to expression (11). In this chapter, the boundary b_k is placed at the point of the maximal probability, as such a decision criterion has appeared to work well in experimental evaluations. This criterion is based on a relative comparison of the evidence of a scene boundary at each point under consideration provided by the feature vector measured at the same point. In this manner, the resulting segmentation procedure resembles the conventional techniques that pronounce scene boundaries at the points of local extremum of some visual or audio similarity curve, expression (11) being considered as a way to fuse multiple data into one cumulative measure. Four posterior probability curves along with audio dissimilarity and video coherence curves obtained for a ground-truth film are depicted in Figure 2. The probability curves are shown partly overlapped; each curve begins at the first candidate point inside a scene, achieves the global maximum at the point of transition to the next scene, and is interrupted at the middle of the next scene (in order not to encumber the figure).

We deliberately include only one local feature vector D_i in expression (11) and exclude surrounding data from consideration. Otherwise, there would be a need to treat properly the strong dependence that usually exists between contiguous observable data. This would complicate the proposed approach and possibly would require more learning data. Experimental tests on a more complicated model that includes the complete set of observable data up to the point under examination, much as the model proposed by Vasconcelos and Lippman (1997) for the task of shot segmentation, suggest that simple neglect of this dependence in such a model degrades considerably the segmentation performance, let alone the increase of the computational complexity. For the same reasons, we do not adopt hidden Markov models that assume conditional independence between observable feature vectors. The problem of dependence between feature vectors is avoided in our model, as the single feature vector D_i in expression (11) usually is placed far enough from boundary b_{k-1} at the most points under examination and, thus, does not depend strongly on the feature vector measured at this boundary.

Final Algorithm

The final segmentation algorithm used in this work is resumed as follows.

Segment an input video into shots and assign candidate points of scene boundaries to be the shot transition moments. Estimate feature vector D_i at each point i.

Place the initial scene boundary b_0 at the beginning of the first scene (which is supposed to be given). Select recursively each subsequent scene boundary b_k based on the position of the previous one b_{k-1} through the following steps:

Calculate the posterior probability of k-th scene boundary at each candidate point i of set $\{b_{k-1} + 1, ..., m_k\}$ according to expression (11) in which m_k is defined by expression (10) and is limited by the last candidate point.

Place the next scene boundary b_k at the point of the highest posterior probability.

If a stopping criterion is fulfilled, exit the algorithm.

The stopping criterion is used mostly to keep inside the narrative part of the input video. In this work, we suppose that the position of the last scene boundary is given and that the stopping criterion is fulfilled when scene boundary b_k appears to be closer in time to the last scene boundary than a predefined threshold value that is approximately equal to the mean scene duration.

Feature Extraction

In this section, we propose visual and audio features that provide evidence of the presence or absence of a video scene boundary and describe the corresponding likelihood estimates required in our sequential segmentation approach.

Video Coherence

Our video coherence feature is derived as a continuous generalization of the conventional graph-based approach (Yeung & Yeo, 1996). As mentioned in the section describing related work, in this approach, visually similar shots first are clustered into equivalence classes and labeled accordingly. Then, a scene transition graph is built, and scene boundaries are claimed at cut edges of the graph. Let's consider the following shot clustering technique. First, a similarity matrix for an input video is built, each element *Sim(i,j)* of which is the value of visual similarity between shots *i* and *j*. Then, each pair of shots that are similar enough (i.e., their similarity is higher then a threshold T_{cl}) is merged into one cluster until the whole matrix is exhausted. This is almost a conventional clustering procedure, except the radius of the clusters is not limited. In practice, we consider the shots that are far apart in time and, hence, are not likely to belong to one scene as nonsimilar and never combine them into one cluster. So, we need to treat only the elements of the similarity matrix located near the main diagonal, which makes the computational burden approximately linear with respect to the duration of the video.

Let's define for each shot *i* the following variable:

$$C^0(i) = \max_{a<i, b\geq i} Sim(a,b). \qquad (13)$$

If this variable is less than the clustering threshold T_{cl}, then, according to the proposed clustering technique, it means that there are no common clusters that combine at least one shot preceding shot *i* with shot *i* or with a shot that follows shot *i*. In this and only in this case, there would be pronounced a scene boundary (at the transition to shot *i*) according to the graph-based segmentation method.

Hence, we can reformulate the conventional graph-based procedure of scene segmentation procedure as searching points on the curve C^0 that fall below the threshold value, scene boundaries being claimed in these points. Alternatively, scene boundaries can be pronounced at the points of local minima of the curve.

In real video, visual similarity between shots within the same scene often is not quite high, especially in action films in which there are many dynamic episodes. Because of this, minima of the variable C^0 often are pronounced badly, and it can happen accidentally that a shot of a scene resembles a shot of the previous or the next scene. In this case, the segmentation procedure can miss scene boundaries. Consider, for example, two scenes represented by a shot clusters chain *ABABCDADCD* in which a real scene boundary occurs before the first shot of cluster *C*, and because of accidental similarity, one of the shots from the second scene was misclassified as *A*. Since the shot clusters in this example cannot be divided into two nonintersecting groups, clustering-based segmenting procedure fails to detect the scene boundary.

In order to enhance the robustness of the segmentation procedure, we can try to implicitly exclude isolated misclassified shots from consideration. At first glance, the next maximal value after C^0 could be taken according to expression (13). However, if a single shot is similar to a shot from another scene, it is likely to resemble other shots of the same cluster. In the previous example of a cluster chain, the shot from the second scene, misclassified

as cluster *A*, is similar to two shots of this cluster for the first scene. Hence, exclusion of a single pair of maximally similar shots does not definitely exclude the influence of a single misclassified shot. So, in addition to this pair, we propose not to take into consideration all the maximally similar shots that follow or precede it and to define for each shot *i* the following variable:

$$C^1(i) = \min\{\max_{a<i,b\geq i,a\neq a_0(i)} Sim(a,b), \max_{a<i,b\geq i,b\neq b_0(i)} Sim(a,b)\}, \quad (14)$$

in which the variables a_0 and b_0 are the shot numbers, for which the expression (13) attains the maximum:

$$\{a_0(i), b_0(i)\} = \arg\max_{a<i,b\geq i} Sim(a,b). \quad (15)$$

By recursion, we can derive variables to exclude the influence of the second misclassified shot, the third one, and so forth:

$$C^n(i) = \min\{\max_{a<i,b\geq i,a\notin\{a_0(i),\ldots,a_{n-1}(i)\}} Sim(a,b), \max_{a<i,b\geq i,b\notin\{b_0(i),\ldots,b_{n-1}(i)\}} Sim(a,b)\}, \quad (16)$$

$$a_n(i) = \arg\max_{a<i,b\geq i,a\notin\{a_0(i),\ldots,a_{n-1}(i)\}} Sim(a,b), \quad (17)$$

$$b_n(i) = \arg\max_{b\geq i,b\notin\{b_0(i),\ldots,b_{n-1}(i)\},a<i} Sim(a,b). \quad (18)$$

The variable C^k has sharp local minima at scene boundaries only if they correspond to *k* misclassified shots. Otherwise, these minima are not well-pronounced. Generally, as the same pair of maximally similar shots can correspond to several contiguous shots, the previously defined variables *C* can remain constant during a period of time. If this constant region corresponds to a local minimum, the scene boundary position cannot be located precisely. In order to use all the variables *C* together and to reduce the probability of wide local minima, an integral variable is defined:

$$C_{\text{int}}(i) = \frac{1}{N}\sum_{k=0}^{N-1} C^k(i), \quad (19)$$

in which *N* denotes the number of terms *C* determined by expression (13) through (18). By analogy with Kender and Yeo (1998), we refer to variable $C_{\text{int}}(i)$ as video coherence and consider it a visual feature that provides evidence of the presence or absence of a scene boundary at the beginning of shot *i*.

The similarity *Sim(a,b)* between shots *a* and *b* involved in expression (13) through (18) can be calculated in various manners. In our experimental evaluations that will be described next, it is calculated as normalized color histogram intersection for the pair of maximally similar key frames representing the shots. The histogram is defined in HSV-color space quantized at 18 hue, 4 saturation, and 3 value points, and included additional 16 shades of

gray. Video coherence feature includes three terms; that is, in expression (19), N is equal to 3 and is calculated for two contiguous groups of five shots that adjoin the point under consideration.

In our sequential segmentation approach, we consider the video coherence feature as a random value generated by a stationary process. In order to evaluate the likelihood of this variable mentioned in expression (1), we use a nonparametrical estimate of the corresponding pdf based on a Gaussian kernel and obtained for a set of presegmented ground-truth data. It is calculated separately for each possible value of the distance to the closest scene boundary Δ. We assume that this distance is limited by a range $[-n_1, n_2]$ in which n1 and n2 are natural numbers of the order of value N in expression (19). If it happens that $\Delta < -n_1$, we set $\Delta = -n_1$, and if $\Delta > n_2$, we set $\Delta = n_2$.

Audio Dissimilarity

In order to calculate the short-term acoustic feature vector for a sound segment, we divide the spectrum obtained from Continuous Wavelet Transform (CWT) into windows by application of triangular weight functions W_i with central frequencies f_i in Mel scale as it is done in the case of Mel Frequency Cepstrum Coefficients calculation (see Figure 3). Unlike the FFT, which provides uniform time resolution, the CWT provides high time resolution and low frequency resolution for high frequencies, and low time resolution with high frequency resolution for low frequencies. In that respect, it is similar to the human ear, which exhibits similar time-frequency resolution characteristics (Tzanetakis, Essl, & Cook, 2001).

Then energy values E_i in each spectral window are computed, and finally, the matrix of spectral bands ratios is obtained as:

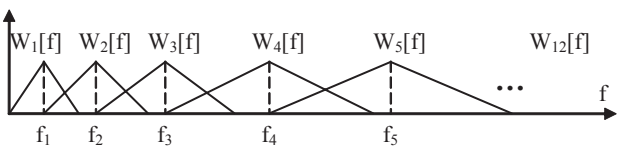
$$K_{ij} = \log(E_i / E_j). \quad (20)$$

Values from the top-right or bottom-left corner of the matrix K are taken as our acoustic features.

The mentioned acoustic feature vector (matrix) is not affected by main volume change, unlike spectral coefficients. At the same time, it allows us to detect changes in acoustic environment.

The procedure of audio dissimilarity curve calculation is done by moving two neighboring windows (with size 8 and step 0.5 seconds in our experiments) along the audio stream and

Figure 3. Triangular weight functions with central frequencies in Mel scale



by obtaining the distance between the distributions of the corresponding acoustic features. Various measures may be used as a distance or dissimilarity for the task of acoustic segmentation: Bayesian Information Criterion (Chen & Gopalakrishnan, 1998), Second-Order Statistics (Bimbot, Magrin-Chagnolleau, & Mathan, 1995), Kullback-Leibler (KL) distance applied directly to distribution of spectral variables (Harb & Chen, 2003).

The KL-measure is a distance between two random distributions (Cover & Thomas, 2003). In the case of Gaussian distribution of random variables, the symmetric KL distance is defined as:

$$K(X_1, X_2) = \left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2}\right) + (\mu_1 - \mu_2)^2 \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right), \tag{21}$$

in which μ and σ are the mean value and the variance of compared distributions.

Instead of multi-dimensional KL applied to a feature vector of spectral bands ratios, a sum of KL distances applied to each element of the vector is used in this work as audio dissimilarity measure:

$$D = \sum_j K(K1_j, K2_j), \tag{22}$$

in which K1 and K2—feature matrices for the neighboring windows.

As an observable feature of a scene boundary in the audio domain, in this work, we extract the maximal value of audio dissimilarity in a time window of about four seconds centered in the corresponding candidate point in order to tolerate small misalignments between the audio and image streams of video. The likelihood of this feature included in expression (4) is calculated from the joint probability as:

$$P(a \mid \tau) = \frac{P(a, \tau)}{P(\tau)}, \tag{23}$$

in which, as earlier, *a* stands for the feature value, τ—for the time distance to the closest scene boundary. We approximate the joint probability with a nonparametric estimate of pdf using a Gaussian kernel on a set of learning data. Just as for the visual feature, we limit the range of τ by a value having the order of duration of the neighboring time windows used to calculate the audio dissimilarity.

Experiments

In this section, we report the results of experiments that are designed to test the proposed video scene segmentation approach. For the lack of common benchmark data, a database of four ground-truth movies (drama: *A Beautiful Mind*; mystery: *Murder in the Mirror*; French

comedy: *Si J'Etais Lui*; and romance: *When Harry Met Sally*) was prepared and manually segmented into semantic scenes. The performance comparisons were made inside time intervals that had the total duration of about 22,000 seconds and included 234 manually labeled scene boundaries.

The segmentation into shots at the preprocessing stage was carried out automatically using a twist-threshold method (Zhang, Qi, & Zhang, 2001) based on color histogram similarity measure. To reduce the computational complexity of the segmentation algorithm, likelihood values of audio and visual features in expressions (1) and (4) were calculated using linear interpolation between tabled values. In the feature domains (fixed through all experiments described next), where estimates of the corresponding pdf became unstable due to the lack of learning data, the likelihoods were extrapolated as constant functions. Only a small portion of data fell into these domains, and experimental evaluations demonstrated that the choice of their boundaries was not crucial for segmentation performance.

Segmentation performance was measured by using the value of precision p and recall r defined as:

$$r = \frac{n_c}{n_c + n_m}, \quad p = \frac{n_c}{n_c + n_f}, \tag{24}$$

in which n_c, n_m, and n_f are the number of correctly detected scene boundaries, the number of missed boundaries, and the number of false alarms, respectively. Detected scene boundary was considered correct if it coincided with a manual scene boundary within an ambiguity of five seconds. Otherwise, it was considered a false alarm. A manual scene boundary was considered missed if it did not coincide with any of the automatically detected boundaries within the same ambiguity of five seconds. As a unified performance score, F1 measure was used:

$$F1 = \frac{2p}{r + p}. \tag{25}$$

Segmentation performance of the proposed sequential segmentation algorithm relative to various films entered into our database is compared in Table 1. Feature likelihoods and scene duration pdf were estimated on the learning set including all four films. The highest integral performance F1 for the film *Murder in the Mirror* was caused mainly by the most stable behavior of the video coherence curve, as the scenes were shot by relatively slow-moving or static cameras. In contrast, the outsider film *Si J'Etais Lui* was characterized by intensive camera movements. A reason for a relatively low performance for the film *A Beautiful Mind* was a less accurate shot segmentation for gradual shot breaks, which sometimes merged shots that were contiguous to a scene boundary.

In order to evaluate the generalization capability of the segmentation approach learned on a set of presegmented data, the cross-validation tests were carried out. The learning set included three films, and the test set consisted of the resting fourth. The overall results for all four films are given in Table 2. Three trials were made: the first one did not use cross-validation at all, serving as a reference; the second used a separate set to learn only the pdf estimates for the audio and visual features, while the scene duration pdf was estimated on a

Table 1. Performance of the sequential segmentation algorithm for various films

Film	Precision, %	Recall, %	F1, %
A Beautiful Mind	67.7	67.7	67.7
Murder in the Mirror	88.9	66.7	76.2
Si J'Etais Lui	66.7	63.2	64.9
When Harry Met Sally	69.8	71.2	70.5
Total for four films	72.4	67.1	69.6

common set including all four films; the third trial supposed separate learning and test sets for all the pdf estimates. As it follows from Table 2, our segmentation approach does not suffer much from parameters over-fitting, providing quite a general model for video scene segmentation. The perceptible sensitivity to the estimate of scene duration pdf suggests the importance of taking into account of prior information about scene duration. The results given next in this section assume the same learning and test set, which includes all four films of the ground truth.

The capability of our sequential segmentation approach to fuse audiovisual features is shown in Table 3, in which the first row presents the segmentation performance when only the visual feature was used, the second row gives the performance only for the audio feature, and the

Table 2. Performance of the sequential segmentation algorithm in cross-validation tests

Using Cross-Validation	Precision, %	Recall, %	F1, %
Non	72.4	67.1	69.6
For the feature pdf only	69.9	67.5	68.7
Total for the feature pdf and the scene duration pdf	67.6	65.0	66.2

Table 3. Performance of the sequential segmentation algorithm for audio-visual feature fusion

Feature Used	Precision, %	Recall, %	F1, %
Visual	61.7	64.1	62.9
Audio	39.9	48.7	43.8
Visual + Audio	72.4	67.1	69.6

third for both features. As it follows from the table, fusing the visual and audio features enhances both recall and precision.

In order to compare our segmentation approach with related multi-modal techniques, we considered the following rule-based scene segmentation algorithm. First, an input video was segmented solely in the visual domain. Strong scene boundaries then were claimed as the actual scene boundaries, while weak scene boundaries were kept only if they were confirmed by a high level of the audio dissimilarity that had to be above threshold *A*. This is a scheme of audiovisual data fusion somewhat analogous to that of Cao, Tavanapong, Kim, and Oh (2003). We also refused from the use of scene duration distribution and adopted a segmentation technique that searched scene boundaries at local minima of the video coherence curve. A local minimum was claimed as a scene boundary if it was a global minimum of enough depth in a surrounding time window and had the absolute value below some threshold *T1*. A scene boundary was considered weak if the corresponding video coherence value was above a second threshold *T2*, *T1>T2*; otherwise, it was marked as a strong boundary. Thresholds *A*, *T1*, and *T2* were selected in order to maximize the overall performance measure F1.

The performance of this rule-based algorithm is given in the first row of Table 4, where it can be compared with the performance of our earlier maximum likelihood ratio approach (Parshin, Paradzinets, & Chen, 2005) and the sequential segmentation one derived in this work. The maximum likelihood ratio algorithm uses the same audio feature as the others; as the visual feature, it used the video coherence C^0 given by expression (13) since it is less dependent on the context and, hence, is more suitable for this algorithm. To compare the efficiency of audiovisual data fusion provided by our rule-based algorithm, we also include the test results for a segmentation algorithm, referenced as "local minima of video coherence," which works solely in the visual domain. This algorithm detects scene boundaries at local minima on the video coherence curve in the same way as our rule-based algorithm with the difference that it uses only one threshold value that maximizes performance measure F1. A comparison of the results given in Table 4 allows us to conclude that the sequential segmentation approach has the best performance measured by both precision and recall.

As for computational time required by our sequential segmentation algorithm, it is quite fast, given that audiovisual features are precomputed and take less than a second on our

Table 4. Performance of different segmentation approaches

Segmentation approach	Precision, %	Recall, %	F1, %
Rule-based	61.0	63.9	62.4
Local minima of video coherence	54.1	64.3	58.8
Maximum likelihood ratio	63.2	63.2	63.2
Sequential segmentation	72.4	67.1	69.6

Intel Pentium 4 1.8GHz computer for one film. This is because the computational complexity is approximately linear with respect to the film length due to a limited time search for each scene boundary. The main computational burden for a raw video file stems from its decoding and feature extraction, which, however, can be done in real time without much optimization for MPEG 4 video format.

Conclusion

A statistical video scene segmentation approach is proposed that combines multiple mid-level features in a symmetrical and flexible manner. In contrast to its rule-based counterparts, it deals properly with real-valued observable features by taking into account the variability of scene boundary evidence provided by these features. This approach also models the duration of scenes, including it as prior information. Two kinds of features are proposed to be used in scene segmentation: video coherence and audio dissimilarity extracted in the visual and the audio domain, respectively. In contrast to the video coherence measure obtained using a conventional short-term memory model, the measure proposed in this work compares only the shots that probably are taken by one camera. Currently, our approach fuses only two types of observable features, but it easily can be extended to include new data. The results of experimental tests carried out on ground truth video showed enhancement of the segmentation performance when multiple modalities are fused. Superior performance also was demonstrated with respect to a rule-based segmentation algorithm.

As our future work, we expect to extend the proposed approach to new features. Useful information, for example, could be provided by automatic person tracking since the same scene usually includes the same personages. New features may appear to be strongly conditionally dependent on each other. So there would be a need to propose more complicated

fusion framework. Also, we are going to apply our approach to other types of video (e.g., sports broadcasting, news programs, or documentary video).

References

Bimbot, F., Magrin-Chagnolleau, I., & Mathan, L. (1995). Second order statistical measures for text-independent speaker identification. *Speech Communication, 17*(1-2), 177–192.

Boresczky, S., & Rowe, L.A. (1996). A comparison of video shot boundary detection techniques. *Proceedings of the SPIE Conference on Storage & Retrieval for Image and Video Databases IV* (pp. 170–179).

Borwell, D., & Thompson, K. (1997). *Film art: An introduction* (5th ed.). New York: McGraw-Hill.

Cao, Y., Tavanapong, W., Kim, K., & Oh, J. (2003). Audio assisted scene segmentation for story browsing. *Proceedings of the International Conference on Image and Video Retrieval* (pp. 446–455).

Chen, S.C., Shyu, M.L., Liao, W., & Zhang, C. (2002). Scene change detection by audio and video clues. *Proceedings of the IEEE ICME* (pp. 365–368).

Chen, S.S., & Gopalakrishnan, P.S. (1998). Speaker environment and channel change detection and clustering via the Bayesian Information Criterion. *Proceedings of the DARPA Speech Recognition Workshop.*

Cover, T., & Thomas, J. (2003). *Elements of information theory*. John Wiley & Sons.

Duda, R.O., & Hart, P.E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Harb, H., & Chen, L. (2003). A query by example music retrieval algorithm. *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS03),* London (pp. 122–128).

Jiang, H., Zhang, H., & Lin, T. (2000). Video segmentation with the support of audio segmentation and classification. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME2000).*

Kender, J.R., & Yeo, B.L. (1998). Video scene segmentation via continuous video coherence. *Proceedings of the IEEE CVPR* (pp. 367–373).

Lienhart, R. (1999). *Comparison of automatic shot boundary detection algorithms*. *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases VII* (pp. 290–301).

Mahdi, W., Ardebilian, M., & Chen, L. (1998). Improving the spatio-temporel clues by the use of rhythm. P*roceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL'98),* Heraklion, Crete, Greece (pp. 169–181).

Mahdi, W., Ardebilian, M., & Chen, L. (2000). Automatic scene segmentation based on exterior and interior shots classification for video browsing. *Proceedings of the IS&T/*

SPIE's 12th International Symposium on Electronic Imaging.

Parshin, V., Paradzinets, A., & Chen, L. (2005) Multimodal data fusion for video scene segmentation. *Proceedings of the 8th International Conference on Visual Information Systems (VIS2005),* Amsterdam, The Netherlands (pp. 279–289).

Rabiner, L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77, no. 2* (pp. 257–286).

Rasheed, Z., & Shah, M. (2003). *A graph theoretic approach for scene detection in produced videos*. Proceedings of the Multimedia Information Retrieval Workshop, Toronto, Canada.

Sundaram, H., & Chang, S.F. (2000). Determining computable scenes in films and their structures using audio-visual memory models. *Proceedings of the ACM Multimedia* (pp. 95–104).

Tzanetakis, G., Essl, G., & Cook, P. (2001). *Audio analysis using the discrete wavelet transform*. Proceedings of the WSES International Conference on Acoustics and Music: Theory and Applications (AMTA 2001), Skiathos, Greece.

Vasconcelos, N., & Lippman, A. (1997). *A Bayesian video modeling framework for shot segmentation and content characterization*. Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries.

Yeung, M.M., & Yeo, B.L. (1996). Time-constrained clustering for segmentation of video into story units. *Proceedings of the International Conference on Pattern Recognition, Vol. C* (pp. 375–380).

Zhang, D., Qi, W., & Zhang, H.J. (2001). A new shot boundary detection algorithm. *Proceedings of the IEEE Pacific Rim Conference on Multimedia* (pp. 63–70).

Section III

Image and Video Annotation