IDIAP RESEARCH REPORT

Finding Structure in Consumer Videos by Probabilistic Hierarchical Clustering

Daniel Gatica-Perez [a] Alexander Loui [b]
Ming-Ting Sun [c]

IDIAP–RR 02-22

May 24, 2002

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, Martigny, Switzerland
[a] Eastman Kodak Company, Rochester, NY, USA
[c] Dept. of Electrical Engineering, University of Washington, Seattle, WA, USA

IDIAP Research Report 02-22

Finding Structure in Consumer Videos by Probabilistic Hierarchical Clustering

Daniel Gatica-Perez Alexander Loui Ming-Ting Sun

May 24, 2002

submitted for publication

Abstract. Accessing, organizing, and manipulating home videos constitutes a technical challenge due to their unrestricted content and their lack of storyline. In this paper, we present a methodology to discover cluster structure in home videos which uses video shots as the unit of organization, and is based on two concepts: (i) the development of statistical models of visual similarity and temporal duration and adjacency of consumer video segments, and (ii) the reformulation of hierarchical clustering (merging) as a sequential binary Bayesian classification process. A Bayesian formulation allows for the incorporation of prior knowledge of the structure of home video, and offers the advantages of a principled methodology. Gaussian mixture models are used to represent the class-conditional distributions of inter-segment visual similarity, and temporal adjacency and duration. The models are then used in the probabilistic clustering algorithm, where the merging order is a variation of Highest Confidence First, and the merging criterion is Maximum a Posteriori. The algorithm does not need any ad-hoc parameter determination. We present extensive simulation results on a ten-hour home video database with ground-truth (30 video sequences, $> 10^6$ frames) which thoroughly validate the performance of our methodology with respect to cluster detection, individual shot-cluster labeling, and the effect of prior selection.

1 Introduction

Among all sources of video content, consumer video probably constitutes the one that most people would eventually be interested in dealing with. The organization and edition of personal memories contained in home videos constitute a technical challenge due to the lack of efficient tools. The development of browsing and retrieval techniques for home video would open doors to the organization of video events in video albums, video baby books, edition of postcards with stills extracted from video, multimedia family web pages, etc. [21], [18], [23].

Unrestricted content and the absence of storyline are the main characteristics of consumer video. Home videos are composed of a set of *events*, each composed of one or a few video shots, visually consistent, and randomly recorded along time. Such features make consumer video unsuitable for analysis approaches based on storyline models [12], and have diverted research on home video analysis until recently, as it was generally assumed that home videos lack of any structure [21], [18], [20], [25], [47], [15]. However, recent studies have revealed that home filmmakers behavior induces certain structure [20], [15], different from that of other video sources [17], [45], as people implicitly follow certain rules of *attention focusing* and *recording*. On one hand, they keep their interest on what they record only for a *limited amount of time*. Additionally, people show their attention by interacting in *specific ways* with the camera. On the other hand, capturing home video imposes *continuity* when recording portions of *the same event*. The structure induced by these filming trends is often semantically meaningful. Based on these observations, recent work has explored the use of motion features to detect significant frames [20]. We further argue that the *cluster structure* of home video can be disclosed from such rules, based on the development of statistical models of visual and temporal features of video segments [1].

In this paper, we propose a novel methodology to discover the cluster structure in home videos, which uses video shots as the unit of organization, and is based on two concepts: (i) the development of statistical models of visual similarity and temporal adjacency and duration of home video segments, and (ii) the reformulation of hierarchical clustering as a sequential binary classification process. A Bayesian formulation allows for the incorporation of prior knowledge of the structure of home video, and offers the advantages of a principled methodology [11], [13]. Our formulation requires the determination of a feature space, and the selection of probability models. Gaussian mixture models (GMMs) are used to represent the class-conditional distributions of the observed features, which are selected based on a study of class separability, in order to minimize the probability of segment misclassification. The models are then used in the hierarchical clustering algorithm, where the merging order is a variation of Highest Confidence First (HCF), and the merging criterion is Maximum a Posteriori (MAP). The algorithm does not need any ad-hoc parameter determination. The cluster structure provides nonlinear access for browsing and manipulation. Our methodology has been rigorously evaluated on an ten-hour (30 video sequences) home video database for which a third party ground-truth is available. On this hard dataset, our methodology produces good performance with respect to cluster detection and individual shot-cluster labeling, validating it as the first step for a system for interactive organization and retrieval of home video information.

The rest of the paper is organized as follows. Section 2 discusses the main features of home video. Section 3 reviews the previous work in home video analysis, and discusses relations with our work. Section 4 presents an analysis of the cluster structure of a home video database, discussing the features that will be exploited by our approach. In Section 5, hierarchical clustering is reformulated

[1] In this paper, the term *cluster* describes the concatenation of one or more scenarios that are filmed in the same physical location (e.g. inside a room, or on a street). Camera motion usually generates several scenarios (several backgrounds) for one single cluster. On the other hand, an *event* is a higher-level semantic entity that is composed of one or several clusters, which might involve more than visual information for its definition (e.g. "living room", "birthday party"). While both *cluster* and *event* convey semantic meaning, the former is usually better defined. A cluster corresponds to an "elementary" event. However, event definition in higher semantic terms is a much more complex task, and outside of the scope of this paper. Furthermore, we reserve the use of the term video *sequence* to denote an entire video file. In contrast, a video *segment* denotes a part of a video sequence composed of one or more shots. In other words, a shot is an "elementary" segment.

as sequential binary Bayesian classification. The selection of the feature space, and the specification and estimation of class-conditional and prior distributions are described in Sections 6 and 7. The performance evaluation and the results of our methodology are discussed in Section 8. Finally, Section 9 draws some concluding remarks.

2 What is Consumer Video?

Several characteristics distinguish home video from other video sources:

- Unrestricted, non-edited content.
- Absence of storyline.
- Temporally ordered information.
- Partially available time-stamp information.
- Frequent poor-quality content (illumination, defocusing).
- Few complex cuts.
- Camera motion. While some types of motion are random (fast panning or hand shaking), others are clearly intentional (e.g., zoom-and-hold) [20].
- Non-continuous audio information. Patterns of the type "short speech/long silence" occur frequently. Additionally, ambient background sound (music, multiple voices) is often found.

The structure of home video bears similarity to the structure of consumer still pictures [23], [33], [31]: videos (resp. film rolls) contain series of *ordered and temporally adjacent shots* (resp. photos) that can be organized in *clusters* that convey semantic meaning. Visual similarity and temporal ordering are indeed two main criteria that allow people to identify clusters in video (resp. picture) collections, when they do not know anything else about the content (unlike the filmmaker or photographer, who knows details of context) [23], [33]. Furthermore, the nature of home video recording generates two special features:

- People can focus their attention when filming only for a *limited amount of time*, which translates both in the amount of time that people can concentrate to record a shot and in the number of shots they film. Previous work has shown that the shot duration in home videos presents patterns [20]. Furthermore, we will show that video clusters themselves also present patterns, in terms of cluster duration, and number of shots per cluster.
- Capturing home video imposes *continuity* when recording. Unlike other video sources [17], [45], [46], filming home video with a temporal back-and-forth structure is rare: on a vacation trip, people do not usually visit the same site twice. In other words, home video content tends to be *localized* in time.

3 Previous Work

Cluster analysis [19] constitutes one of the challenges of video analysis [48], [46], [16], [34], [18], [15]. In particular, hierarchical agglomerative clustering (HAC) methods have been used in the past [46], [22], [15], [47]. Early work [48], [46], [34] proposed versions of either purely visual or time-constrained shot clustering, without specifically addressing home video. Specifically, the work in [46] proposed the use of a time-constrained full-link HAC, which required the heuristic set up of a number of thresholds for visual and temporal features, and the used of a transition graph for browsing. The methodology was applied for structuring of TV programs and movies.

In the home video domain, the work in [21] used shot clustering to extract automatic video summaries, assuming time-stamped materials. This methodology relied on time-and-date detection, and clustering using arbitrary temporal thresholds to find clusters at different time scales, without using any further visual information. The works in [18] and [20] were the first ones that explicitely analyzed some of the inherent statistics of home video. The work in [18] created multiple video segment groupings to provide different views of the content, based on probabilistic feature descriptions. After representing each video shot by a discrete distribution, video shots were clustered by a soft top-down, annealing method based on information-theoretic cost criteria. The work in [20] presented an analysis on three hours of consumer video that revealed *patterns* both in shot duration and camera motion. A heuristic algorithm to detect zoom-and-hold motion was then used to extract *significant* frames, and compose video summaries without recurring to shot detection. The work in [25] described a system for analysis of home video based on the detection and tracking of faces inside video shots.

As described in the previous section, home videos and consumer pictures bear similarity with each other. In this direction, [47] recently presented an extension of a clustering methodology for consumer pictures [31] to videos. The work in [44] aims at classifying vacation still images into a small number of pre-defined classes with semantic meaning (city-landscape, indoor-outdoor) using low-level visual features.

Our work shares the Bayesian methodology with a number of recent approaches for tasks other than video structuring, like [45] (shot boundary detection), and [44] (still image classification). In our case, we want to disclose the cluster structure of videos for which the number of classes and their meaning cannot be pre-defined. A pre-established number of classes could be used for clustering, but would be quite limited for unrestricted content. Specifically to video structuring, our work is related to the work in [46], but it is distinct in several ways. Unlike [46], our work systematically investigates visual and temporal features of a specific source of video, and learns probability models that are used for clustering based on an optimality criterion. The probabilistic formulation avoids the use of heuristics that are hard to define, and allows to model multiple features in a unified way (a joint distribution), and to introduce additional knowledge using a prior distribution.

4 Analyzing the Cluster Structure of Consumer Video

4.1 The Kodak Consumer Video Database

Our data set consists of 30 MPEG-1 video clips of different characteristics. Each sequence has a duration between 18 and 25 minutes, and was digitized from VHS tapes at 1.5Mb/s in SIF format. The total duration is nearly ten hours (about 1075000 frames). The data were collected from eleven different people, and is representative of consumer video content: indoor and outdoor scenarios, depicting weddings, vacations, children playing, school parties, etc. A third-party ground-truth at both the shot and the cluster levels was manually generated (see Section 8 for further discussion). Additionally, garbage video shots (transitional shots with no content), and very poor quality shots were not taken into account for the experiments. These two situations mimic human handling of poor quality still pictures. After this adjustment, the total number of shots and clusters in the database are 801 and 189, respectively. Both the number of shots and the number of clusters per sequence vary considerably (between 4 and 105 shots, and between 1 and 19 clusters, respectively). As far as we are aware, this constitutes the most extensive home video database reported in the current literature.

4.2 Analyzing the Cluster Structure in Home Videos

4.2.1 The Effect of Limited Focus-Of-Attention

Statistical modeling of temporal video features was originally proposed in [45], introducing a Weibull model for shot duration in professional movie trailers. A similar approach was followed in [20], modeling home video shot duration with log-normal distributions. While in the first case shot duration is closely related to the creation of narrative atmospheres [17], [45], in the second one it constitutes an expression of human interest. However, this feature was not used in [20], as the authors claimed that shot duration did not appear to be related to the frame significance.

We argue that not only shots, but also home video clusters have clear temporal patterns. Unlike [20], we have made use of this information. Let $\tau \in [0, \infty)$ denote shot duration. Fig. 1(a) illustrates its empirical distribution, and an approximation by a GMM [26],

$$p(\tau|\mathcal{I}) = \sum_{i=1}^{N} \omega_i p(\tau|\theta_i, \mathcal{I}), \tag{1}$$

where N is the number of components in the mixture, ω_i denotes the prior probability of the i-th component, $p(\tau|\theta_i, \mathcal{I}) = \mathcal{N}(\mu_i, \sigma_i)$ is the i-th univariate Gaussian distribution parameterized by $\theta_i = \{\mu_i, \sigma_i\}$, and $\mathcal{I}$ is the prior information assumed about the world [13]. It can be seen that the length of home video shots mostly remains in the range of a couple of minutes. This is a clear indication of the typical amount of time that people are able to keep interested when operating a camera. Furthermore, this limitation in interest is also evident when measuring the duration of video clusters. Let $\Gamma \in [0, \infty)$ denote home video cluster duration. Fig. 1(b) shows the empirical distribution of cluster duration and its mixture approximation, $p(\Gamma|\mathcal{I})$. Video clusters have a definite trend to last only a few minutes. In our database, approximately 95% of the clusters last less than ten minutes. As a consequence of lack of attention, very long video clusters are rare.

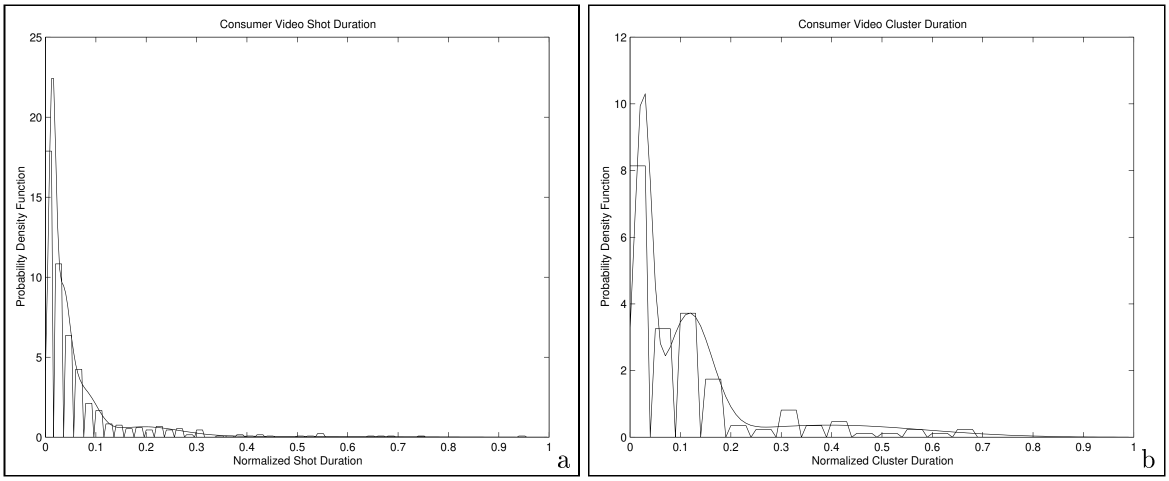

Figure 1: (a) Empirical distribution of normalized shot duration, and its GMM approximation. Shot duration was normalized by the longest shot in the database (580 s.). (b) Empirical distribution of normalized cluster duration, and its GMM approximation. The maximum cluster duration is 1217 s.

The complementary information is the distribution of number of shots per video cluster, denoted by $p(\kappa|\mathcal{I})$. Although both the number of shots and the number of clusters per sequence (denoted by υ and Υ, respectively) vary considerably, most clusters are composed of only a few shots. Figs. 2(a-b) show the distribution of shots ($p(\upsilon|\mathcal{I})$) and clusters ($p(\Upsilon|\mathcal{I})$) per sequence. One can see their wide variability (as a general trend, outdoor shots are usually shorter than indoor shots, hence outdoor video sequences normally contain more shots than indoor sequences of similar duration; however, it is also common to find both outdoor and indoor shots in the same video sequence). Fig. 2(c) shows the distribution of number of shots per cluster, $p(\kappa|\mathcal{I})$. In brief, approximately half of the clusters in the database are composed of one or two shots, and four out of five clusters are composed of six or less shots.

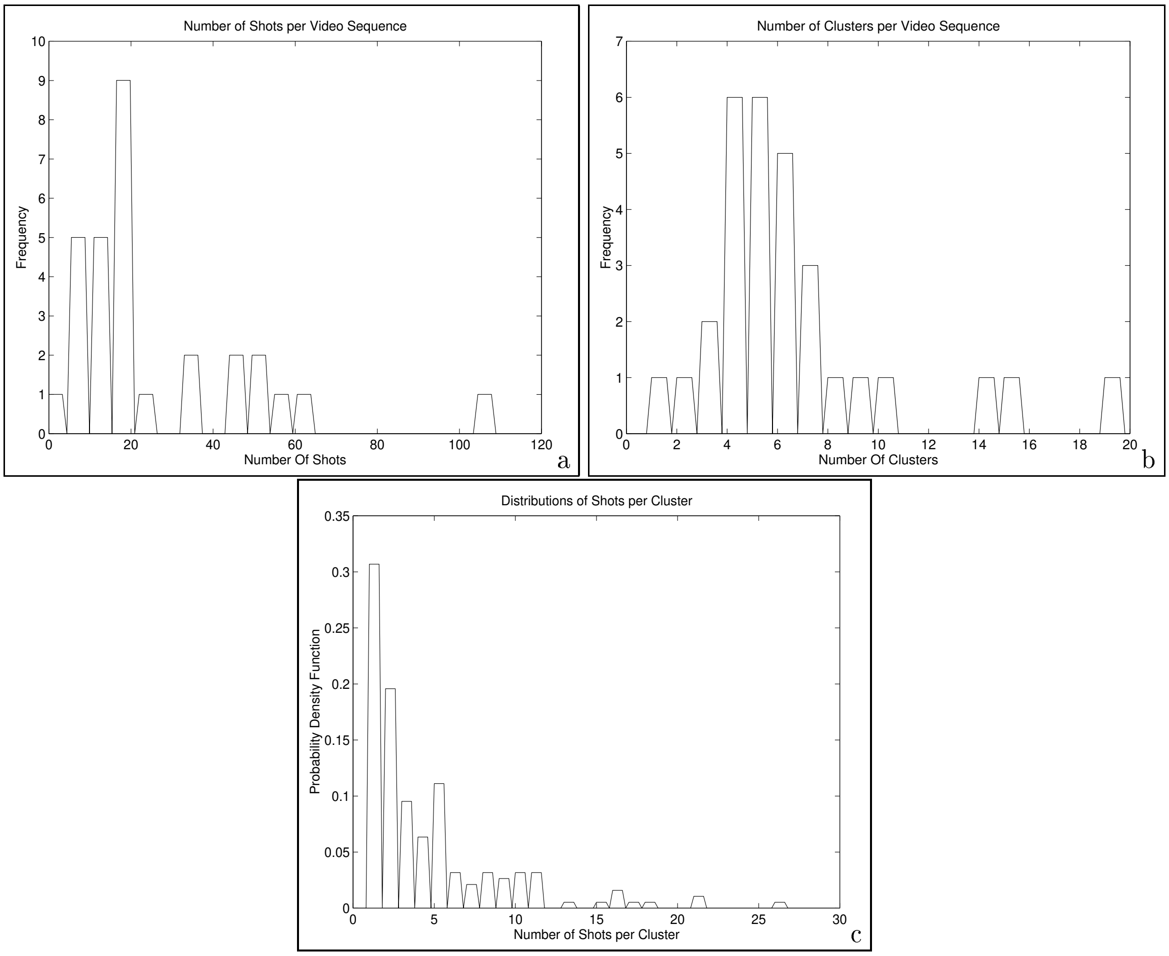

Figure 2: Empirical distributions of (a) number of shots per sequence; (b) number of clusters per sequence; (c) number of shots per cluster. 50.3% of the clusters in the database are composed of one or two shots; 80.4% of the clusters are composed of six or less shots.

4.2.2 The Effect of Continuity

Home video clusters composed of non-adjacent shots (forward and backward temporal jumps) are infrequent. In our database, only about 3% of the clusters present this characteristic. In other words, clusters are *localized* in time, and therefore strong connectivity conditions can be applied for clustering, which has the benefits of computational simplicity.

4.2.3 Visual Similarity in Home Video Clusters

Content-based image analysis has addressed the issue of computing similarity between images/videos [24], [30], [36], [9], [39], [44], [5], [49]. In our problem, a question is that of the visual structure of home video clusters: how similar (resp. dissimilar) are segments that belong to the same (resp. a different) cluster? Let s_i and s_j denote the i-th and j-th segments in a sequence, and let $\mathcal{E}$ be a binary r.v. that indicates their belonging to the same cluster Ω,

$$\mathcal{E} = \begin{cases} 1 & \text{if } \Omega(s_i) = \Omega(s_j) \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

Additionally, let $d(s_i, s_j)$ denote the similarity measure between two segments. We computed a B-bin mean RGB color histogram $h_i = \{h_{iz}\} \in \mathcal{R}^B$ for each shot in the database. Then, we empirically constructed the distributions of intra-cluster ($p(d|\mathcal{E} = 1, \mathcal{I})$) and inter-cluster visual similarity ($p(d|\mathcal{E} = 0, \mathcal{I})$) using pairs of shots. For the inter-cluster case, pair-wise computation was limited within the interval that contains 95% of the probability mass of cluster duration $p(\Gamma|\mathcal{I})$ (Fig. 1). The similarity measure was the norm in the L_1 space $d_{L1}(\cdot)$ [49],

$$d(s_i, s_j) = d_{L_1}(h_i, h_j) = \sum_{z=1}^{B} |h_{iz} - h_{jz}|. \tag{3}$$

The distributions, displayed in Fig. 3 [38], appeared quite overlapped as a result of the unrestricted content of home video. This highlights the limitations of both features and distance measures to define similarity among video segments, and the technical challenge of the problem at hand.

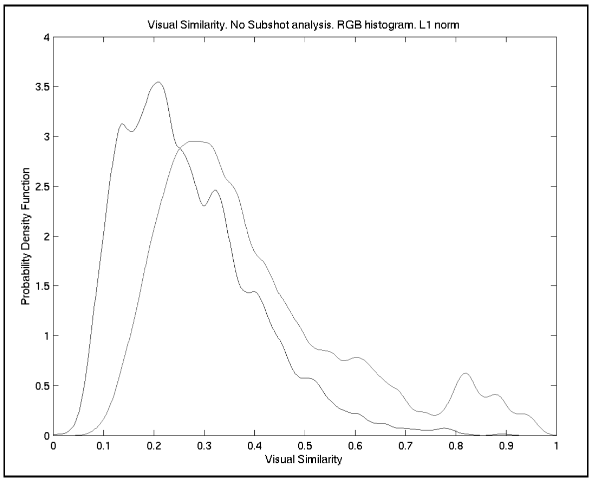

Figure 3: Visual similarity in home video clusters, based on global segment information. Intra-cluster ($p(d|\mathcal{E} = 1, \mathcal{I})$) and inter-cluster ($p(d|\mathcal{E} = 0, \mathcal{I})$) distributions are represented by continuous and dotted-line curves, respectively.

Recapitulating, the analysis of the database shows that there indeed exists cluster structure in home videos. Furthermore, it suggests the development of methods that (i) integrate segment visual similarity and duration in a joint model, (ii) rely on strong temporal adjacency, and (iii) account for the fact that clusters are likely to contain only a few elements. These observations motivated the methodology described in the next section.

5 Our Approach: Probabilistic Hierarchical Clustering of Home Video Segments

HAC algorithms can be based on *probability models* [28], [3], [4]. To capture the inherent characteristics of home video clusters, we propose to build statistical models of visual similarity and temporal duration and adjacency defined on pairs of segments in a HAC framework. In particular, a HAC algorithm can be thought of as a *sequential binary classifier*, which at each step decides whether a pair of segments should be merged according to a probability model. The formulation of clustering as a two-class pattern classification problem allows for the use of Bayesian decision theory [11]. The MAP criterion establishes that given an n-dimensional realization x_{ij} of X (representing features extracted from segments s_i and s_j), the class $\mathcal{E}$ that must be selected is

$$\mathcal{E}^* = \arg\max_{\mathcal{E}} \Pr(\mathcal{E}|x, \mathcal{I}), \tag{4}$$

where $\mathcal{E}$ is defined in Eq. 2, $\Pr(\mathcal{E}|x, \mathcal{I})$ denotes the posterior probability of $\mathcal{E}$ given x, and the subindices in x have been dropped. Applying Bayes' rule and reorganizing terms,

$$L = \frac{p(x|\mathcal{E}=1, \mathcal{I}) \Pr(\mathcal{E}=1|\mathcal{I})}{p(x|\mathcal{E}=0, \mathcal{I}) \Pr(\mathcal{E}=0|\mathcal{I})} \underset{H_0}{\overset{H_1}{\gtrless}} 1, \tag{5}$$

where $p(x|\mathcal{E}, \mathcal{I})$ are the class-conditional pdfs of the observed features given $\mathcal{E}$, $\Pr(\mathcal{E}|\mathcal{I})$ are the priors of $\mathcal{E}$ given the knowledge about the world, L denotes the posterior odds ratio, H_1 denotes the hypothesis that the pair of segments belong to the same cluster, and H_0 denotes the opposite. The priors allow for the introduction of knowledge about the characteristics of home video. Our algorithm starts by treating each elementary segment (shot) as a cluster, successively evaluates the pair of segments that correspond to the largest L, merges only when $L \geq 1$, and continues until H_1 in Eq. 5 is no longer valid for any pair of segments. This greedy strategy bears similarity with the Highest Confidence First (HCF) method used in Bayesian image analysis [7], and is intuitively appealing: at each step, decisions should be made based on the piece of information that has the highest certainty. Additionally, the formulation does not require any ad-hoc parameter determination, and can be seen as a generalization of previous time-constrained clustering algorithms [46]. Due to the characteristics of home video, only the two neighbors of each segment have to be analyzed. The method can be efficiently implemented using adjacency graphs and priority queues [7], [27], [22], as described in the Appendix.

Our methodology, summarized in Fig. 4, requires the determination of a useful feature space, and the selection of models for the distributions. These issues are described in the next sections.

6 Video Segment Feature Extraction and Selection

To generate the basic segments, shot boundaries are detected by a series of standard methods [14]. Oversegmentation due to illumination or noise artifacts can be well handled by the clustering algorithm. In the following subsections, we describe the process of feature extraction and selection, which is based on an empirical study of discriminative power of features and similarity measures in consumer video segments.

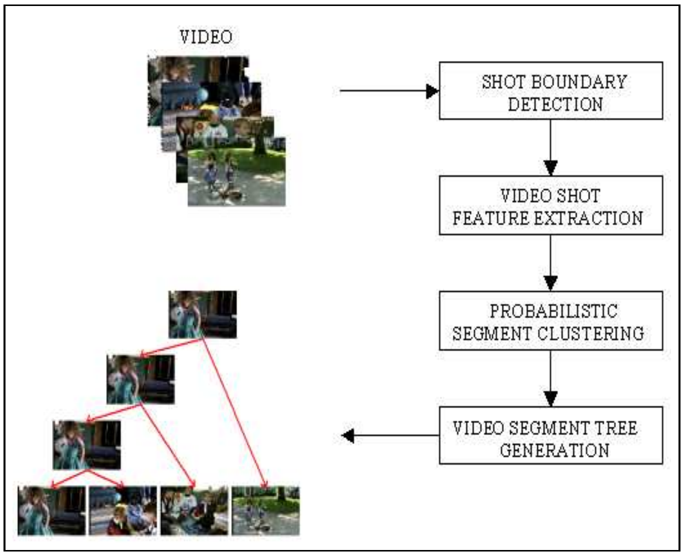

Figure 4: Architecture for Consumer Video Structuring.

6.1 Extraction of Visual Features

Home video shots usually contain more than one appearance, due to hand-held camera motion. Consequently, we have adopted an approach that detects clusters inside each shot (named *subshots* in the following), which approximately correspond to an individual appearance of the scene, and then extracts features from a set of *random* frames in each subshot. This generates a three-level hierarchy as shown in Fig. 5. We consider that sophisticated key-frame extraction algorithms would not outperform random frame selection unless there is theoretical support for it. A shot s_i is defined as a collection of K subshots, $s_i = \{s_{ik}, k \in \{1, ..., K\}\}$, and each subshot is characterized by a set of M random frames, $s_{ik} = \{s_{ikm}, m \in \{1, ..., M\}\}$.

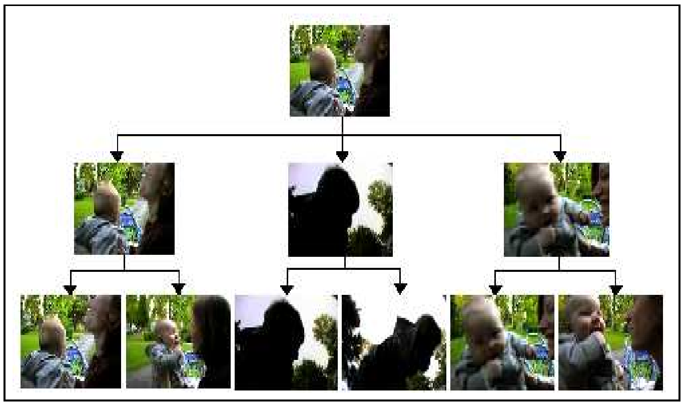

Figure 5: Shot hierarchy. The root node is the shot, the intermediate level is composed of subshots, and the leaves are random frames extracted from each subshot.

In the first place, subshots are sequentially detected using a technique similar to that described in [34] for shot boundary detection. Furthermore, subshots of very short duration are discarded, as they often correspond to fast camera pannings. In the second place, from the various image representations that have been proposed for content-based retrieval [29], [48], [43], [24], [30], [49], [39], [44], we have selected *joint histograms* [30]. These are simple estimations of multivariate distributions that combine color and scene structure information, and have shown to significantly improve color-only image retrieval. Investigated features included:

1. Color in the RGB space (uniformly quantized to $8 \times 8 \times 8$ bins), and the HSV space (vector-quantized to 1024 colors [44]).

2. Color ratios (known to be illumination-invariant), non linearly quantized to 32 levels [1].

3. Edge-based features, including edge density and edge directions [43].

Other features can be tested and added in a straightforward fashion [29].

The final point consists in the definition of the similarity measure. If the subshots s_{ik}, s_{jl} are characterized by M and N random frames respectively, each represented by a joint histogram h_{ikm}, h_{jln}, the similarity between two subshots is defined as

$$d(s_{ik}, s_{jl}) = \min\{d_{\phi}(h_{ikm}, h_{jln}), m \in \{1, ..., M\}, n \in \{1, ..., N\}\}. \tag{6}$$

where other measures, like the metric based on the Bhattacharyya coefficient d_{BT} [8], or a measure based on the correlation coefficient d_{CC} [40], were substituted in d_{ϕ}. The similarity between two shots s_i and s_j, consisting of K and L subshots respectively, can then be computed as a R-ranked vector of similarities between the subshots in increasing order,

$$d(s_i, s_j) = \{d^r(s_{ik}, s_{jl}), k \in \{1, ..., K\}, l \in \{1, ..., L\}, r \in \{1, ..., R\}\}, \tag{7}$$

where the index r indicates the rank. For $R = 1$,

$$d(s_i, s_j) = \min\{d(s_{ik}, s_{jl}), \forall k \in \{1, ..., K\}, l \in \{1, ..., L\}\}. \tag{8}$$

Alternatively, probabilistic measures of similarity could be defined if the dimensionality of the feature space were reduced [2].

6.2 Selection of Visual Features

We estimated the distributions $p(d|\mathcal{E} = 0, \mathcal{I})$ and $p(d|\mathcal{E} = 0, \mathcal{I})$ for all the features and similarity measures discussed in the previous subsection, following the procedure described in Section 4.2, Eq. 8, and a subset of 75% of the sequences in our database. The most discriminative features were selected based on the overlap between the two distributions. The empirical probability of error, assume noninformative priors for class selection, can be computed by

$$\Pr(e|\mathcal{I}) = \frac{1}{2}(\Pr(e|\mathcal{E} = 0, \mathcal{I}) + \Pr(e|\mathcal{E} = 1, \mathcal{I})), \tag{9}$$

where $\Pr(e|\mathcal{E} = 0, \mathcal{I})$ and $\Pr(e|\mathcal{E} = 1, \mathcal{I})$ are the overlapped areas between the two class-conditional pdfs of visual similarity measures. Tables 1-2, and Fig. 6 summarize the results.

The advantage of using subshot detection and random frames (SS+RF) as opposed to global shot information can be seen by comparing Fig. 3 and Fig. 6(a). The subshot analysis has increased the separability between the two classes. The use of joint histograms further improve discrimination. Table 1 shows the empirical probability of error computed for RGB histograms with and without subshot detection, and for 4D histograms that combine color and edge density (EDEN), edge directions (EDIR), and color ratios on the Y component (YR). No significant improvement was found when using the HSV color model. From the tested options, RGB-EDEN produced slightly better results than the other 4D histograms. Additionally, the results of applying different metrics are presented in Table 2 and Fig. 6(b-d). The L_1 norm and the Bhattacharyya metric produced better results than the correlation coefficient which is not a good measure for this purpose. The Bhattacharyya coefficient can be interpreted as the cosine of the angle between the component-wise square-rooted pdfs approximated from the corresponding joint histograms [8], so d_{L_1} and d_B can be thought of as representations of both magnitude and angle, and constitute the features to characterize visual similarity in our method.

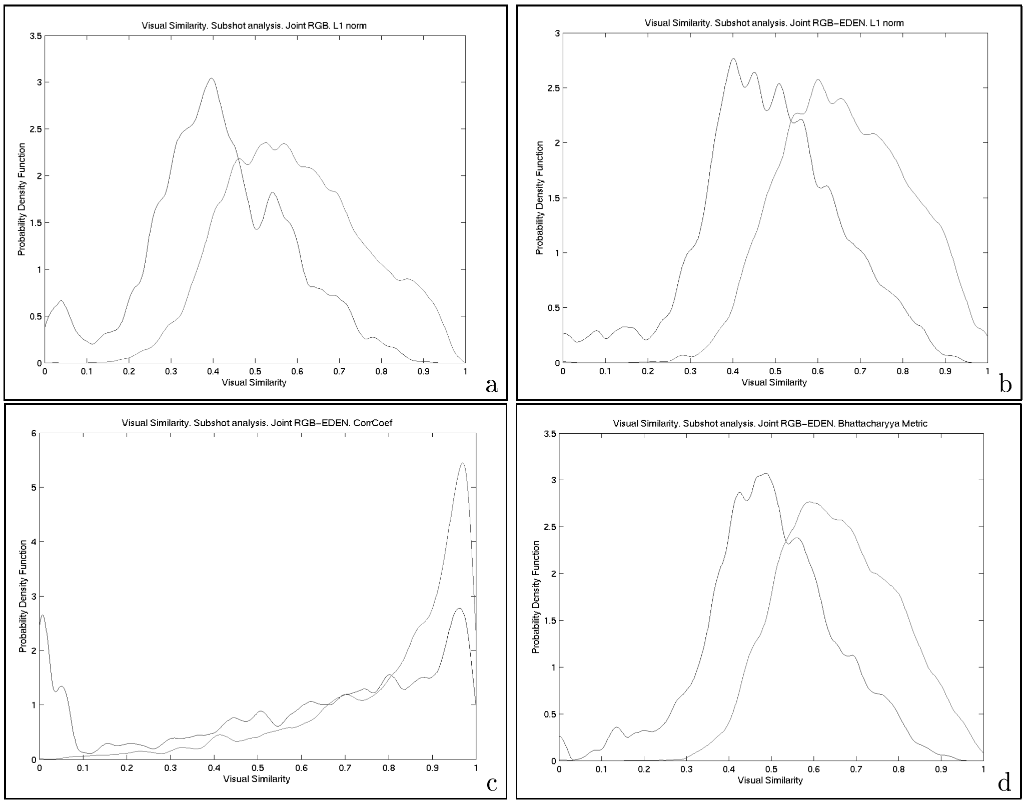

Figure 6: Visual similarity features with subshot detection and random frame extraction. $p(d|\mathcal{E} = 1, \mathcal{I})$ and $p(d|\mathcal{E} = 0, \mathcal{I})$ are represented by continuous and dotted-line curves, respectively. (a) RGB histograms, L_1 norm. (b-d) Joint RGB-edge density histograms with (b) L_1 norm, (c) correlation coefficient, (d) metric based on the Bhattacharyya coefficient.

Joint Hist	Type	Dimension	Pr(e)
RGB	Shots Only	512	0.364
RGB	SS + RF	512	0.319
RGB-YR	SS + RF	16384	0.295
RGB-EDIR	SS + RF	3584	0.286
HSV-EDIR	SS + RF	7168	0.284
RGB-EDEN	SS + RF	5120	0.280

Table 1: Feature Selection. L_1 metric

Measure	Pr(e)
d_{L_1}	0.280
d_{CC}	0.365
d_B	0.292

Table 2: Comparison of Similarity Measures. Joint Histogram RGB-EDEN

6.3 Selection of Temporal Features

The analysis in Section 4 made evident the possibility of using strong adjacency for clustering. The accumulated length of two individual segments is an indication about their belonging to the same cluster (segments of increasing length will become less likely to belong to the same video cluster). The accumulated segment duration is defined as

$$\Delta_{ij} = min\{|e_j - b_i|, |e_i - b_j|\} \tag{10}$$

where b_i and e_i denote the first and last frame of segment s_i. This definition provides a joint probabilistic formulation for similar features used in previous work [46], [34].

7 Data Analysis: Modeling of Likelihood Functions and Prior

7.1 Modeling of Likelihood Functions with GMMs

The described features constitute the feature space, with vectors $X = (d_{L1}, d_{BT}, \Delta) \in \mathcal{X}$. Fig. 7(a) shows a scattering plot of 600 vectors from our database, projected onto the first and third components of $\mathcal{X}$. While visual similarity itself is clearly limited to perform clustering, the accumulated segment length is a good feature to classify pairs of segments.

The joint class-conditional pdfs of the observed features are represented by multivariate GMMs,

$$p(x|\mathcal{E}, \Theta, \mathcal{I}) = \sum_{i=1}^{N_{\mathcal{E}}} \omega_i p(x|\mathcal{E}, \theta_i, \mathcal{I}), \tag{11}$$

where $N_{\mathcal{E}}$ is the number of components in each mixture, ω_i denotes the prior probability of the i-th component, $p(x|\mathcal{E}, \theta_i, \mathcal{I}) = \mathcal{N}(\mu_i, \Sigma_i)$ is the i-th d-dimensional Gaussian with full covariance matrix, parameterized by $\theta_i = \{\mu_i, \Sigma_i\}$,

$$p(x|\mathcal{E}, \theta_i, \mathcal{I}) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \tag{12}$$

and $\Theta = \{\{\omega_i\}, \{\theta_i\}\}$ represents the set of all parameters. [26]. The Expectation-Maximization (EM) algorithm constitutes the standard procedure for Maximum Likelihood (ML) parameter estimation for a broad range of problems where the observed data are in some sense incomplete [10]. In the

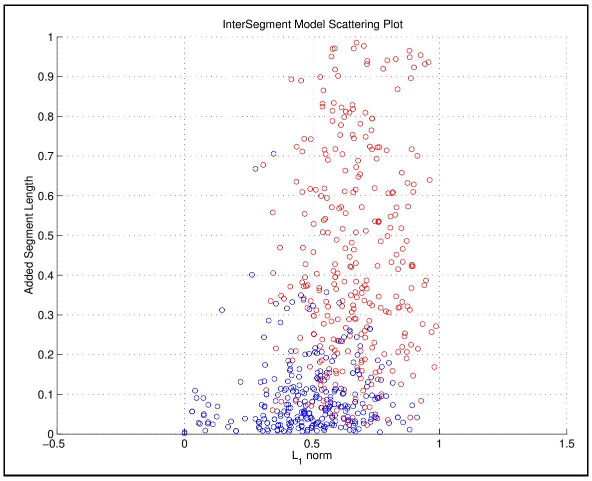

Figure 7: Projection onto the first and third components for 300 randomly selected training vectors from each class. Intra-cluster and inter-cluster feature vectors are represented by blue and red circles, respectively.

case of a GMM, the incomplete data are the unobserved mixture components, whose prior probabilities are the parameters ω_i. Additionally, model selection can be automatically estimated using the Minimum Description Length (MDL) principle [32], by choosing

$$N_{\mathcal{E}}^* = \arg\max_{N_{\mathcal{E}}} (\log L(\Theta|\mathcal{X}) - \frac{n_{N_{\mathcal{E}}}}{2} \log N), \tag{13}$$

where $L(\cdot)$ denotes the likelihood of the training set, N is the number of training vectors, $\mathcal{X}$ is the training set, and $n_{N_{\mathcal{E}}}$ is the number of parameters needed for the model, given by

$$n_{N_{\mathcal{E}}} = (N_{\mathcal{E}} - 1) + N_{\mathcal{E}} d + N_{\mathcal{E}} \frac{d(d+1)}{2}. \tag{14}$$

When two models fit the sample data in a similar way, the simpler model is chosen.

7.2 Modeling of Prior

In the Bayesian approach, the prior $\Pr(\mathcal{E}|\mathcal{I})$ encodes the belief about the merging process [13]. The simplest assumption is $\Pr(\mathcal{E} = 0|\mathcal{I}) = \Pr(\mathcal{E} = 1|\mathcal{I})$. However, the study in Section 4 suggested that merging must be discouraged as many clusters consist of only a few shots. The prior should reflect this knowledge. One possibility is to determine it from the available evidence, i.e., to estimate it from training data. While in strict terms, this technique does not conform to the Bayesian principles, it usually produces better solutions than arbitrary priors [11]. Assuming independence in the N training data, the ML estimator of the prior is defined by

$$\Pr(\mathcal{E} = \epsilon|\mathcal{I}) = \frac{1}{N} \sum_{k=1}^{N} i(\epsilon, k) \tag{15}$$

where $i(\epsilon, k) = 1$ if the k-th training sample belongs to the class $\mathcal{E} = \epsilon$, and zero otherwise.

8 Experiments and Results

When cluster structure does exist in data (so that a ground-truth can be generated), the two criteria for quantitative evaluation of a clustering algorithm are $\mathcal{C}_1$: the determination of the number of clusters, and $\mathcal{C}_2$: the determination of the cluster label for each datum, compared to the ground-truth [11]. Although many algorithms for video segment clustering have been proposed [48], [46], [16], [6], [34], [21], [22], their performance using the two mentioned criteria is unknown in several cases. Furthermore, we are not aware of any comparative study of video segment clustering techniques. Table 3 compares the performance evaluation procedure of some commonly referenced algorithms, as extracted from the original references. $\mathcal{E}_1$ indicates if the reference reported the detected number of clusters, and $\mathcal{C}_1$ and $\mathcal{C}_2$ indicate whether the two above criteria were used for evaluation. $\mathcal{NA}$ indicates that some piece of information was not available.

Reference	Clips	Type of Video	Frames	Shots	$\mathcal{E}_1$	$\mathcal{C}_1$	$\mathcal{C}_2$
[48]	1	Documentary	18000	71	Yes	No	No
[46]	5	Movies/TV	132051	1027	Yes	No	No
[6]	23	TV	> 400000	$\mathcal{NA}$	No	No	No
[16]	4	Movies	$\mathcal{NA}$	$\mathcal{NA}$	Yes	Yes	Yes
[34]	7	Movies	176554	1396	Yes	Yes	No
This paper	30	Home	1075348	801	Yes	Yes	Yes

Table 3: Performance Evaluation Procedure of Segment Clustering Algorithms

8.1 Ground-truth

In our database, a third-party cluster ground-truth was determined based on human evaluation of shot visual similarity, temporal adjacency, and *blind* context understanding [23]. The use of third-party ground-truth is common for performance evaluation, including movie analysis [16] and still images [37], and is useful to perform benchmarking against the limit of a computer algorithm which has no context knowledge, thus producing a fairer estimate of its performance.

Note that, although there are differences of judgement between different people due to the uncertainty about the contents, there indeed exists cluster structure in home videos. The incorporation of human similarity judgements for evaluation of video analysis algorithms represents a valuable research area. Some related work has been done for still images databases [41], [33], but there is virtually no work done in the home video domain. An evaluation methodology currently under development consists in the definition of statistical measures of human judgment in a Bayesian context, and their use to validate computer algorithms.

8.2 Performance Evaluation Procedure

Results were generated with the leave-one-out method: one video sequence was held for evaluation while all the remaining were included in the training set for density estimation [11]. Given NC, the number of clusters in the ground-truth (either in an individual sequence or in the whole database), the criterion $\mathcal{C}_1$ is evaluated by defining three variables: *Detected Clusters* (DC), which indicates the number of clusters that were found by the algorithm, *False Positives* (FP), defined by $FP = DC - NC$ if $DC - NC \geq 0$ and zero otherwise, and *False Negatives* (FN), defined by $FN = NC - DC$ if $DC - NC \leq 0$ and zero otherwise. To evaluate the criterion $\mathcal{C}_2$, *Shots In Error* (SIE) is used to denote the number of shots whose cluster label does not match the label in the ground-truth. Finally, *Correcting Operations* (CO) indicates the number of operations (merging/splitting) needed to correct the results so that SIE is zero. We believe this is a good indication of the effort required in interactive systems. We are interested in analyzing the performance figures as probabilities. If z is any of the parameters of interest, the frequentist evaluation of the performance produces two typical estimates: the *macro-average* z_M, which is directly computed over the whole database, and the *micro-average* z_m, in which the figure is first estimated for each individual sequence, and then averaged over the whole database. While the former assigns the same importance to each shot (or cluster) in the database, the latter gives the same importance to each video sequence, regardless of the number of shots or clusters it contains. We present both figures for discussion. For macro-averages, if $FP = 0$, the figures are computed by

$$dc = \frac{DC}{NC}; \quad fp = 0; \quad fn = \frac{FN}{NC}, \tag{16}$$

and if $FN = 0$, the expressions are

$$dc = \frac{DC - FP}{DC}; \quad fp = \frac{FP}{DC}; \quad fn = 0. \tag{17}$$

Additionally,

$$sie = \frac{SIE}{NS}; \quad co = \frac{CO}{NS}, \tag{18}$$

where NS stands for the number of shots. For micro-averages, Eqs. 16-18 are valid for each individual sequence. Results are then accumulated and averaged over the whole database.

8.3 Results

The detailed results are shown in Table 4; the summarized results appear in Tables 5 and 6. Table 5 evaluates the capability of our methodology to detect clusters. This is a very hard problem, due to the variability in the number of clusters per sequence in home video, as highlighted in Fig. 2(b). The macroaverage shows that the total number of detected clusters approximately corresponds to the total number of clusters in the database. This is obviously an over-optimistic estimate as false positives in some sequences compensate for false negatives in others when computing the sample mean. In contrast, the micro-average constitutes a more reliable performance measurement for cluster detection. The estimated value for dc was 0.75 (the ground-truth would produce a value of one). Furthermore, fp is approximately twice the value of fn (0.171 and 0.079, respectively), which refects the fact that our method has a tendency to oversegmentation (from Table 4, the algorithm generated at least one false positive in 16 sequences, and at least one false negative in 8 sequences). A similar trend has been reported by other researchers for other types of video content [46], [34]. Furthermore, a detailed observation of the results indicated that several of the false negatives (not detected clusters) actually consist only of one or two shots according to the the ground-truth. As a baseline, we show the poor result that is obtained with an algorithm that randomly estimates the number of clusters for each video in the database. This result simulates the case in which home videos truly did not have structure, so any random clustering would be equally good.

Table 6 describes the performance of our algorithm in terms of shot-cluster assignment. For macro- and micro-averages, the ground-truth generates a zero value for sie and co. In this case, both the macro-average and micro-averages are useful measurements. Variations between them indicate difference of performance from sequence to sequence. We have selected a number of baseline methods for comparison, which assume the *correct* number of clusters for each sequence in the database. The methods are (i) B_1, which assigns a uniform and temporally adjacent number of shots per cluster [31]; (ii) B_2, a version of K-means clustering for shots, in which the centroids were initialized with randomly selected shots from each sequence; and (iii) B_3, the same variation of K-means, but in which the centroids were initialized with equally "spaced" shots (in terms of shot number) extracted from each sequence. The distance between a shot and a centroid in the K-means algorithm was computed by using Eq. 8. The shot representation (random frames extracted from subshots, each represented by a 4-D joint histogram) remained constant for all clustering algorithms. Finally, we also included B_0, the case of random clustering.

Video-clip	Duration	Shots	Clusters	DC	FP	FN	SIE	CO
V_0	20:01	15	4	4	0	0	3	2
V_1	21:17	19	2	3	1	0	9	1
V_2	18:21	18	8	6	0	2	3	3
V_3	20:12	19	5	6	1	0	6	3
V_4	21:39	18	6	10	4	0	4	4
V_5	19:47	18	5	5	0	0	3	3
V_6	20:01	47	14	13	0	1	17	9
V_7	20:01	59	15	13	0	2	23	10
V_8	20:01	54	10	15	5	0	21	10
V_9	22:45	62	7	12	5	0	18	7
V_{10}	20:01	35	6	5	0	1	9	2
V_{11}	18:52	48	7	6	0	1	20	5
V_{12}	23:01	8	3	7	4	0	4	4
V_{13}	25:01	11	4	7	3	0	5	4
V_{14}	20:01	7	4	5	1	0	1	1
V_{15}	20:02	10	5	7	2	0	2	2
V_{16}	20:01	8	3	1	0	2	3	1
V_{17}	20:01	19	5	7	2	0	4	3
V_{18}	18:02	4	1	4	3	0	3	3
V_{19}	23:57	18	4	5	1	0	9	5
V_{20}	23:07	105	19	7	0	12	28	6
V_{21}	20:49	10	4	5	1	0	2	2
V_{22}	19:56	12	4	5	1	0	2	1
V_{23}	20:27	35	9	9	0	0	6	3
V_{24}	20:00	18	5	6	1	0	4	2
V_{25}	20:00	12	7	7	0	0	0	0
V_{26}	20:00	54	6	4	0	2	10	3
V_{27}	20:00	16	6	6	0	0	3	2
V_{28}	20:00	20	5	6	1	0	5	3
V_{29}	20:00	22	6	6	0	0	3	2
Total	617:34	801	189	202	36	23	230	106

Table 4: Video Clustering Results on Kodak Consumer Video Database.

The results show that our methodology outperformed all of the baseline methods, even with the assumption of using the correct number of clusters. Using macro-averages (resp. micro-averages) as measurement, our methodology assigned 71.1 (resp. 71.4)% ($100 * (1 - sie)$) of the shots to the correct cluster. In contrast, the two K-means algorithms performed approximately the same, the best one generating 47.6 (resp. 56)% of correct shot assigments. Interestingly, uniform shot-assigment performed better than K-means (54.7 (resp. 57)% of correct assigments). A similar trend can be observed for the probability of correcting operations (co). The mean number of shots per sequence is $801/30 = 26.7$, and therefore with our method 3.55 (resp. 4.62) operations are needed in average to correct the cluster assignments in a 20-minute video.

We can use the Bayesian approach to specify a prior on the probability of shot in error, include a likelihood, and use the posterior (conditioned on the observations) to compute posterior intervals or visualize the performance [13]. In our N-shot database, suppose we observe n shots in error. The likelihood function $\Pr(n|sie)$ is a binomial distribution, $\Pr(n|sie) \propto sie^n(1-sie)^{N-n}$. If, for analytic convenience, we further assume a uniform prior $p(sie)$, the expression for the posterior after applying

Method	dc_M	fp_M	fn_M	dc_m	fp_m	fn_m
Random Clustering	0.305	0.655	0.000	0.470	0.514	0.015
Probabilistic Clustering	0.934	0.065	0.000	0.750	0.171	0.079

Table 5: Cluster Detection Performance

Method	sie_M	co_M	sie_m	co_m
B_0	0.679	0.609	0.588	0.529
B_1	0.453	0.167	0.430	0.200
B_2	0.533	0.407	0.462	0.373
B_3	0.524	0.398	0.440	0.348
Probabilistic Clustering	0.289	0.133	0.286	0.173

Table 6: Shot Assignment Performance

Bayes' rule is

$$p(sie|n) \propto sie^n (1 - sie)^{N-n}.$$

Fig. 8(a) compares the posterior distributions over the probability of shot in error, estimated for the different clustering methods, where $N = 801$ (the distributions have been rescaled in the vertical axis to be plotted together). Fig. 8(b) presents the corresponding analysis to compare the posterior distributions of the probability of correcting operations $p(co|n)$.

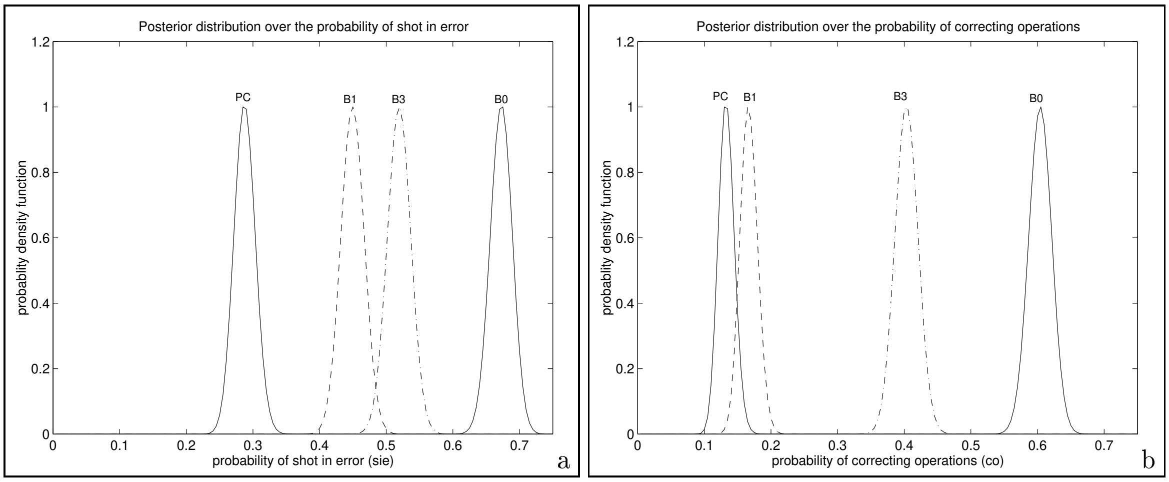

Figure 8: (a) Posterior distributions of the probability of shot in error, assuming a uniform prior, for different structuring algorithms. PC denotes our approach. (c) Posterior distributions of the probability of correcting operations.

The effect of the prior distribution is shown in Table 7. The use of a uniform prior does not make use of knowledge of the problem: merging should be discouraged as most video clusters consist only of a few number of shots. The results reflect this fact: no false positives were detected in the entire database, as more mergings were allowed, but this additional clustering resulted in performance detriment. The uniform prior generates a micro-average $dc = 0.573$, and a probability of shot in error of 0.393 and 0.309, using macro- and micro-averages, respectively. A detailed inspection of the results indicate that larger clusters have indeed been favored, and most errors come from shots that belong to clusters which contained one or two shots and which were erroneouly merged. On the other hand, the estimation of the prior based on Eq. 15 for our database produced $\Pr(\mathcal{E}|\mathcal{I}) = \{0.87, 0.13\}$. This distribution better modeled the knowledge about the problem, and indeed improved performance.

Method	dc_M	fp_M	fn_M	sie_M	dc_m	fp_m	fn_m	sie_m
Uniform prior	0.470	0.000	0.529	0.393	0.573	0.000	0.427	0.309
Empirical prior	0.934	0.065	0.000	0.289	0.750	0.171	0.079	0.286

Table 7: Effect of Prior Probability

Two examples of the generated clusters are shown in Figs. 9 and 10. Each cluster is displayed as a row of shots, which are in turn represented by a random frame each. Qualitatively, our methodology provides quite reasonable results. We have used the proposed methodology as the first step in the development of a system for organization of consumer video. A prototype of an interface that displays the structure as a tree [35] consisting of frames extracted from the video, cluster, shot, and subshot levels, is shown in Fig. 11. The interface allows for the manipulation of the video structure, including retrieval of basic information about the structure, and correction of it. Playback capabilities based on the video structure for efficient browsing are under development.

8.4 Limitations

Fig. 12 shows typical merging errors. The images were extracted from pairs of adjacent shots that were erroneously merged by our algorithm. As a general trend, outdoor clusters are harder to segment correctly. There are three main reasons for erroneous merging: (1) high visual similarity between semantically disjoint but temporally adjacent video clusters, as discussed in Section 6.2, (2) shots of very short duration, and (3) clusters of very short duration. The two reasons for erroneous oversegmentation of clusters are (1) very high intra-cluster visual variability, and (2) unusually long clusters.

Although we have shown that our approach produces good results in a real, hard dataset, it is known that the use of global low-level features has limitations in modeling semantic information, including the human definition of visual similarity [49], [36], [39]. Our work could benefit from the use of image segmentation schemes in order to improve the measurements of visual similarity among video segments. Image segmentation into a few regions or blobs could be used as the starting point for matching elements across images. Similar approaches have been useful for retrieval of still images [39]. The applicability of such an approach in home video remains to be seen. One advantage of our approach is that the definition of new features (including for instance multiple definitions of similarity) can be directly introduced in the formulation, via a joint pdf. Finally, while we have not used higher-level features such as faces, we are considering the additional computation cost of analyzing images to extract these features.

9 Concluding Remarks

This paper proposed a systematic methodology to discover cluster structure in consumer video, by incorporating the inherent characteristics of such content in a probabilistic framework. A detailed analysis of the visual and temporal structure of a relatively large and diverse database offered a number of clues that were embedded in a Bayesian formulation of hierarchical clustering. Home video features of intra and inter-cluster visual similarity, adjacency, duration and composition were exploited. The results obtained by our methodology are encouraging, but also illustrate the complexity of the research problem.

Several issues remain open. As discussed earlier, the investigation of mechanisms to quantify similarity between video segments, and of features and algorithms that can capture such definitions are challenging unsolved issues. The investigation of region-based and multimedia representations

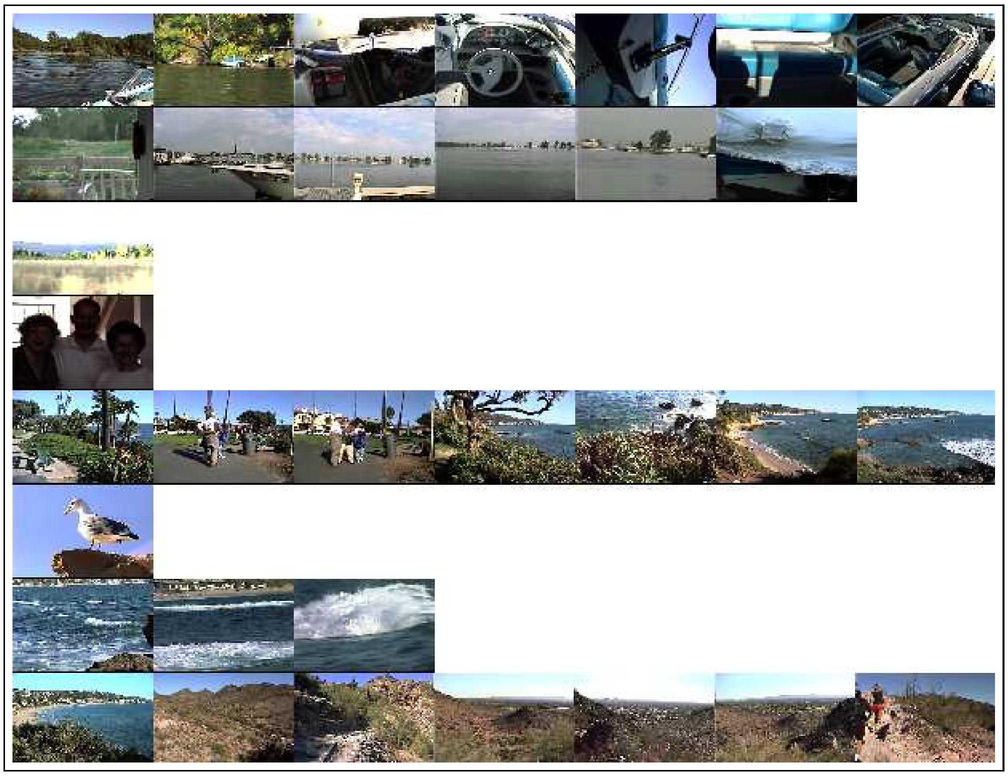

Figure 9: Generated structure on a *Vacation* sequence (detail). Each shot is represented by one frame.

Figure 10: Generated structure on a *Baby* video sequence (detail). Each shot is represented by one frame.

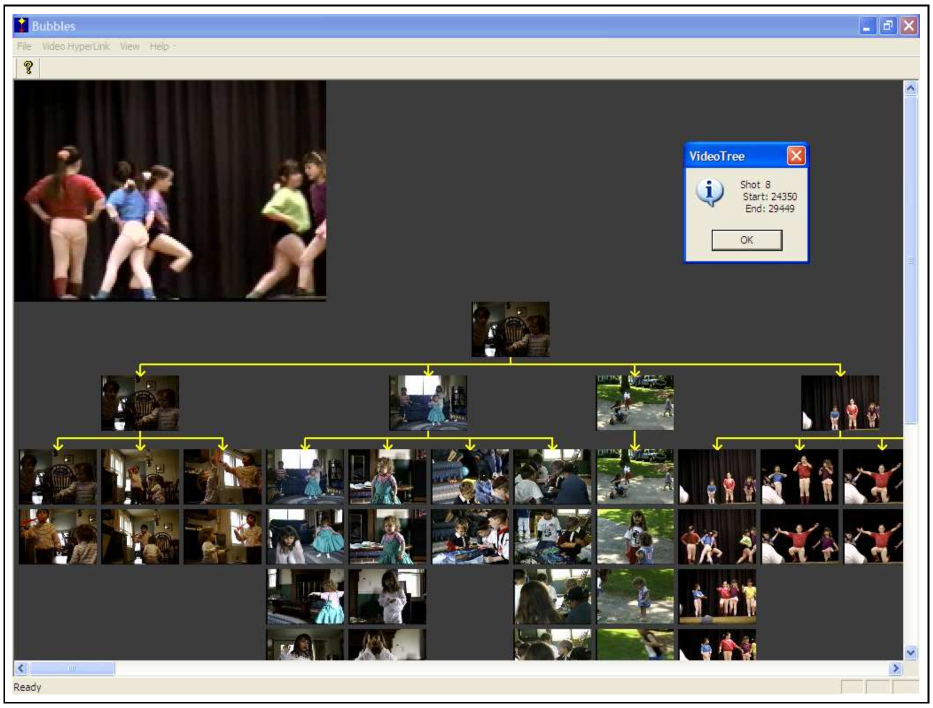

Figure 11: Displaying the Video Segment Tree.

Figure 12: (a-b), (c-d) Frames extracted from pairs of video shots that were erroneously merged by our methodology.

(i,e., adding audio features), and its integration in a probabilistic framework constitute current lines of work.

10 Appendix

The method consists of two stages: queue initialization and queue updating/depletion.

1. *Queue initialization.* At the beginning of the process, all segments are composed of one shot. Inter-segment features x are computed, and introduced in the queue with priority equal to the binary likelihood ratio (Eq. 5).

2. *Queue depletion/updating.* Until the queue is empty,

   (a) Extract an element from the queue. This pair of segments is the one that has the highest priority.

   (b) Apply the MAP criterion to merge the pair of segments (Eq. 5)

   (c) If the segments are merged (hypothesis H_1), update the model of the merged segment, then update the queue based on the new model, and go to step 2. Otherwise (H_0), go to step 2.

When a pair of segments is merged, the model of the new segment s'_i is updated by

$$\begin{aligned} s'_i &= \{s'_{io}, o \in \{1, ..., K+L\}\} \\ &= \{s_{ik}, k \in \{1, ..., K\}\} \cup \{s_{kl}, l \in \{1, ..., L\}\} \\ b'_i &= \min(b_i, b_j) \\ e'_i &= \max(e_i, e_j) \\ card(s'_i) &= card(s_i) + card(s_j) \end{aligned} \tag{19}$$

Additionally, a subset of the random frames that represent each segment are taken out, and new random frames are introduced. After having updated the model of the (new) merged segment, four functions are implemented to update the queue: (i) extraction from the queue of all those elements that involved the originally individual (now merged) segments, (ii) computation of new inter-segment features using the updated model, (iii) computation of new priorities, (iv) insertion in the queue of elements according to new priorities.

Acknowledgments

The authors thank Peter Stubler for providing software for shot boundary detection, Salvador Ruiz-Correa for valuable discussions, Napat Triroj for helping with the database collection, and the Eastman Kodak Company for providing the Home Video Database.

References

[1] D. A. Adjeroh, and M. C. Lee, “On Ratio-Based Color Indexing,” *IEEE Trans. on Image Processing*, Vol. 10, No. 1, pp. 36-48, Jan. 2001.

[2] S. Aksoy and R. M. Haralick, “Probabilistic vs. Geometric Similarity Measures for Image Retrieval,” in *Proc. IEEE Conf. on Comp. Vis. and Patt. Rec.*, Hilton Head Island, S.C., June 2000.

[3] J. D. Banfield and A. E. Raftery, “Model-based Gaussian and Non-Gaussian Clustering,” *Biometrics*, Vol. 49, No. 3, pp. 803-821, Sept. 1993.

[4] H. H. Bock, “Probabilistic Models in Cluster Analysis,” *Computational Statistics and Data Analysis*, Vol. 23, pp. 5-28, 1996.

[5] R. Brunelli, O. Mich, and C.M. Modena, “A Survey on the Automatic Indexing of Video Data,” *Journal of Visual Communication and Image Representation*, Vol. 10, pp. 78-112, 1999.

[6] J-Y Chen, C. Taskiran, A. Albiol, E. J. Delp and C. A. Bouman, “ViBE: A Video Indexing and Browsing Environment,” in *Proc. SPIE Conf. on Multimedia Storage and Archiving Systems IV*, SPIE vol. 3846, Boston, Sept 1999, pp. 148-164.

[7] P. Chou and C. Brown, “The Theory and Practice of Bayesian Image Labeling,” *International Journal of Computer Vision*, Vol. 4, pp. 185-210, 1990.

[8] D. Comaniciu, V. Ramesh, and P. Meer, “Real-Time Tracking of Non-Rigid Objects using Mean Shift,” in *Proc. IEEE Conf. on Comp. Vis. and Patt. Rec.*, Hilton Head Island, S.C., June 2000.

[9] I.J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, “The Bayesian Image Retrieval System,PicHunter: Theory, Implementation, and Psychophysical Experiments,” *IEEE Trans. on Image Processing*, Vol. 9, No. 1, pp. 20-37, Jan. 2000.

[10] A.P. Dempster, N.M Laird, and D.B. Rubin, “Maximum Likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, Series B, 39:1-38, 1977.

[11] R. Duda, P. Hart, and D. Stork. *Pattern Classification*, Second Edition. John Wiley and Sons, 2000.

[12] S. Eickeler and S. Muller, “Content-based Video Indexing of TV Broadcast News Using HMMs," in *Proc. IEEE ICASSP 99*, Phoenix, pp. 2997-3000, 1999.

[13] A. Gelman, J. B. Carlin, H.S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall-CRC, 1996.

[14] U. Gargi, R.Kasturi, and S.H. Strayer, “Performance Characterization of Video-Shot-Change Detection Methods,” , *IEEE Trans. on Circ. and Syst. for Video Tech.*, Vol. 10, No. 1, February 2000, pp. 1-13.

[15] D. Gatica-Perez, M.-T. Sun, and A. Loui, “Consumer Video Structuring by Probabilistic Merging of Video Segments,” in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Tokyo, Aug. 2001.

[16] A. Hanjalic and H.J. Zhang, “An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis,” *IEEE Trans. on Circuits and Systems For Video Technology*, Vol. 9, No. 8, pp. 1280-1289. Dec. 1999,

[17] J. Hart, *Film Directing Shot by Shot : Visualizing from Concept to Screen.* Focal Press, 1991.

[18] G. Iyengar, and A. Lippman, “Content-based browsing and edition of unstructured video,” in *Proc. Int. Conf. on Multimedia and Expo*, New York City, Aug. 2000.

[19] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, New York, 1998.

[20] J.R. Kender and Yeo, B.L., "On the Structure and Analysis of Home Videos," in *Proc. Asian Conf. on Computer Vision*, Taipei, Jan. 2000.

[21] R. Lienhart, "Abstracting Home Video Automatically," in *Proc. ACM Multimedia Conf.*, Orlando, Oct. 1999. pp. 37-41.

[22] J. Llach and P. Salembier, "Analysis of Video Sequences: Table of contents and index creation," in *Proc. Int. Workshop on Very Low Bitrate Video*, Kyoto, Oct. 1999.

[23] A. Loui and M. Wood, "A software system for automatic albuming of consumer pictures," in *Proc. ACM Multimedia 99*, Orlando, Nov. 1999.

[24] W.-Y. Ma and H.J. Zhang, "Content-based Image Indexing and Retrieval," In B. Fuhrt, Ed., *Handbook of Multimedia Computing*, CRC Press, Boca Raton, 1999, pp. 227-254.

[25] W.-Y. Ma and H.J. Zhang, "An Indexing and Browsing System for Home Video," In *Proc. EUSIPCO, European Conference on Signal Processing.* Patras, Greece, 2000, pp. 131-134.

[26] G.J. MacLachlan and D. Peel. *Finite Mixture Models.* John Wiley and Sons, N.Y., 2000.

[27] M. Meila and D. Heckerman, "An Experimental Comparison of Several Clustering and Initialization Methods," Microsoft Research Technical Report MSR-TR-98-06 , Feb. 1998.

[28] S. M. Omohundro, "Best-first model merging for dynamic learning and recognition," in *Proc. Advances in Neural Information Processing Systems,* Vol. 4, pp. 958-969, 1992.

[29] G. Pass, R. Zabih, and J. Miller, "Comparing Images Using Color Coherence Vectors," in *4th ACM Conference on Multimedia*, Boston, MA, Nov. 1996.

[30] G. Pass and R. Zabih, "Comparing Images Using Joint Histograms," *ACM Journal of Multimedia Systems*, 7(3), pp. 234-240, May 1999.

[31] J. Platt "AutoAlbum: Clustering Digital Photographs using Probablisitic Model Merging," in *Proc. IEEE Workshop on Content-Based Access to Image and Video Libraries*, Hilton Head Island, S.C.,2000.

[32] J. Rissanen, "Modeling by shortest data description," *Automatica*, 14: 465-471, 1978.

[33] K. Rodden, "How do people organise their photographs?," in *Proc. BCS IRSG 21st Ann. Colloq. on Info. Retrieval Research*, 1999.

[34] Y. Rui and T. Huang, "A Unified Framework for Video Browsing and Retrieval," in Alan Bovik, Ed., *Image and Video Processing Handbook*, Academic Press, 2000, pp.705-715.

[35] P. Salembier, L. Garrido, "Binary Partition Tree as an Efficient Representation for Image Processing, Segmentation, and Information Retrieval," *IEEE Trans. on Image Processing*, 9(4):561-576, April 2000.

[36] S. Santini and R. Jain, "Similarity Measures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 9, pp. 871-883, Sep. 1999.

[37] A. Savakis, S. Etz, and A. Loui, "Evaluation of image appeal in consumer photography," in *Proc. SPIE/IST Conf. on Human Vision and Electronic Imaging V*, San Jose, CA, Jan. 2000.

[38] D.W. Scott, *Multivariate Density Estimation*, New York, Wiley, 1992.

[39] 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. on Pattern Analysis And Machine Intelligence*, Vol. 22, No. 12, pp. 1349-1380, Dec. 2000.

[40] G.W. Snedecor, and W. G. Cochran, *Statistical Methods,* 7th Edition. The Iowa State University Press, 1980.

[41] D. M. Squire and T. Pun, "Assessing Agreement Between Human and Machine Clusterings of Image Databases," *Pattern Recognition*, 31, 12, pp. 1905-1919, 1998.

[42] A. Stolcke and S. Omohundro, "Hidden Markov Model Induction by Bayesian Model Averaging," in *Proc. Advances in Neural Information Processing Systems,* Vol. 5, pp. 11-18, 1993.

[43] A. Vailaya, A. Jain, and H.J. Zhang, "On Image Classification: City Images vs. Landscapes," *Pattern Recognition*, Vol. 31, pp 1921-1936, Dec. 1998.

[44] A. Vailaya, M. Figueiredo, A. Jain, and H.J. Zhang, "Image Classification for Content-based Indexing," *IEEE Trans. on Image Processing*, Vol. 10, No. 1, pp. 117-130, Jan. 2001.

[45] N. Vasconcelos and A. Lippman, "A Bayesian Video Modeling Framework for Shot Segmentation and Content Characterization," in *Proc. IEEE Comp. Vis. Patt. Rec.* San Juan, 1997.

[46] M. Yeung, B.L. Yeo, and B. Liu, "Segmentation of Video by Clustering and Graph Analysis," *Computer Vision and Image Understanding*, Vol. 71, No. 1, pp. 94-109, July 1998.

[47] L. Zhao, W. Qi. Y.J. Wang, S.Q. Yang, and H.J. Zhang. "Video Shot Grouping Using Best-First Model Merging," in *Proc. Storage and Retrieval for Media Databases*, SPIE vol. 4315, pp.262-269. Jan. 2001.

[48] D. Zhong and H. J. Zhang, "Clustering Methods for Video Browsing and Annotation,' in *Proc. Storage and Retrieval for Still Images and Video Databases IV*, SPIE vol. 2670, pp.239-246, Feb. 1996.

[49] H.J. Zhang, "Content-based Video Browsing and Retrieval," In B. Fuhrt, Ed., *Handbook of Multimedia Computing*, CRC Press, Boca Raton, 1999, pp. 255-280.