Finding Structure in Home Videos by Probabilistic Hierarchical Clustering

Daniel Gatica-Perez, Alexander Loui, and Ming-Ting Sun

***Abstract*—Accessing, organizing, and manipulating home videos present technical challenges due to their unrestricted content and 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: 1) the development of statistical models of visual similarity, duration, and temporal adjacency of consumer video segments and 2) the reformulation of hierarchical clustering 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 intra- and inter-segment visual and temporal features. 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 results on a 10-h home-video database with ground truth which thoroughly validate the performance of our methodology with respect to cluster detection, individual shot-cluster labeling, and the effect of prior selection.**

***Index Terms*—Bayesian decision theory, clustering, home-video structuring.**

I. Introduction

AMONG ALL sources of video content, home video probably constitutes the one that most people would eventually be interested in dealing with. However, the organization and edition of personal memories contained in home videos present technical challenges due to the lack of efficient tools. The development of such tools could open doors to video albuming and other multimedia applications [14], [11], [15].

Unrestricted content and the absence of storyline are the main characteristics of home video. A typical home video contains 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, and have diverted research on home-video analysis until recently, as it was generally assumed that home videos lack of any structure [14], [11]. However, recent studies have revealed that home filmmakers' behavior induces certain structure [13], [7], different from that of other video sources [10], as people implicitly follow rules of attention focusing and recording. On one hand, people keep their interest on what they film only for a limited amount of time, and display their interest 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 investigated the extraction of significant frames [13]. 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 methodology to discover the cluster structure in home videos based on two concepts: 1) the development of statistical models of visual similarity, duration, and temporal adjacency of video segments and 2) the reformulation of hierarchical clustering as a sequential binary classification process. 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. The models are then used in the hierarchical clustering algorithm, where the merging order is a variation of highest confidence first (HCF) [3], and the merging criterion is maximum *a posteriori* (MAP) [9]. The algorithm does not need any *ad-hoc* parameter determination. Our methodology has been evaluated on a 10-h database (30 video sequences) for which a third-party ground truth is available, showing good performance with respect to cluster detection and individual shot-cluster labeling. The cluster structure provides nonlinear video access and can be used in a system for video browsing and retrieval.

The paper is organized as follows. Section II discusses the main features of home video. Section III reviews previous work. Section IV presents an analysis of the cluster structure of home video, discussing the features exploited by our approach. Section V introduces our methodology. The selection of the feature space and probability models are described in Sections VI and VII. The results are presented and discussed in Section VIII. Finally, Section IX draws some concluding remarks.

II. What Is Home Video?

Several characteristics distinguish home video from other video sources:

- unrestricted, nonedited content;
- absence of storyline;

[1]In this paper, the term *cluster* describes the concatenation of scenarios that are filmed in the same physical location (e.g., inside a room). Camera motion usually generates several scenarios for each cluster. On the other hand, an *event* is a higher-level semantic entity composed of one or several clusters, which might involve more than visual information for its definition (e.g., "living room," "birthday party"). Both *cluster* and *event* convey semantic meaning, clusters corresponding to "elementary" events. However, event definition in higher semantic terms is a much more complex task, and outside of the scope of this paper. Furthermore, the term video *sequence* denotes an entire video file. In contrast, a *segment* is a part of a sequence composed of one or more shots; shots are "elementary" segments.

Manuscript received June 9, 2001; revised December 1, 2002. This paper was recommended by Associate Editor H. J. Zhang.

D. Gatica-Perez is with the Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH-1920 Martigny, Switzerland.

A. Loui is with Eastman Kodak Company, Rochester, NY 14650 USA.

M.-T. Sun is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA.

Digital Object Identifier 10.1109/TCSVT.2003.813428

- temporally ordered information;
- partially available time-stamp information;
- frequent poor-quality content (illumination, defocusing);
- few complex cuts;
- camera motion: some patterns are random (hand shaking), but others clearly intentional (zoom-and-hold) [13];
- noncontinuous audio: "short speech/long silence" and ambient background sound are dominant patterns.

The structure of home videos bears similarity to the structure of home still pictures [15], [18]: videos (respectively, film rolls) contain series of ordered and temporally adjacent shots (respectively, photos) that can be organized in clusters that convey semantic meaning. Visual similarity and temporal ordering are indeed two of the criteria that allow people to identify clusters in video (respectively, picture) collections, when they do not know anything else about the content (unlike the filmmaker or photographer, who knows details of context) [15]. Furthermore, home video is characterized by two special features.

1) People can focus their attention when filming only for a limited amount of time. This translates both into the amount of time that people use to record individual shots, and into the number of shots they film per event. Previous work has shown that home-video shot duration presents patterns [13]. Here, we show that video clusters also display patterns in terms of cluster duration and number of shots per cluster.
2) Home-video recording imposes temporal continuity. Unlike other video sources [10], [22], [23], 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, the content tends to be localized in time.

III. Previous Work

Clustering [12] is one of the goals of video analysis [25], [23], [19], [11], [7]. Hierarchical agglomerative clustering (HAC) methods have been used in the past [23], [7], [24]. Early work also proposed visual-based and time-constrained clustering, without specifically addressing home video [25], [23], [19].

Previous work on home-video analysis can be summarized as follows. The work in [14] used shot clustering for video summarization, assuming time-stamped materials and using only temporal information. The works in [11] and [13] were the first ones to explicitly analyze some of the statistics of home video. The first approach created multiple groupings to provide different views of the content, using probabilistic feature descriptions and an information-theoretic-based annealing method [11]. The second one presented an analysis of patterns in shot duration and camera motion, and proposed a heuristic algorithm to extract frames based on detection of zoom-and-hold motion [13]. The work in [16] described a system based on detection and tracking of faces inside video shots. Finally, the work in [24] extended a clustering method developed for consumer pictures [18] to videos.

Our work shares the Bayesian methodology with a number of recent approaches for tasks other than video structuring, like shot-boundary detection [22] and still-image classification [21]. In our case, we want to disclose the structure of videos for which the number of classes cannot be pre-defined. Our work is also related to the work in [23], but it is distinct in several ways. Unlike [23], we systematically investigate visual and temporal features of a specific video source, and use probability models for clustering. Our formulation avoids the use of heuristics that are hard to define, allows to model multiple features in a unified fashion (a joint distribution), and provides a principled way to introduce knowledge about the problem (a prior distribution).

IV. On the Cluster Structure of Home Videos

A. The Kodak Home-Video Database

The data set consists of 30 MPEG-1 video clips, with individual duration between 18 and 25 min, and digitized from VHS tapes at 1.5 Mbit/s in SIF format. The total duration is nearly 10 h. The videos were collected from 11 people, and are representative of consumer content: indoor and outdoor scenes depicting weddings, vacations, children at home, school parties, etc. A third-party ground truth at both the shot and the cluster levels was manually generated (see Section VIII for further discussion). Additionally, transitional shots with no content, and very poor quality shots were not taken into account. After this adjustment, the set consists of 801 shots and 189 clusters. The number of shots and clusters per sequence presents significant variations.

B. Analyzing the Cluster Structure of Home Videos

1) The Effect of Limited Focus of Attention: Statistical models of temporal video features were originally proposed in [22], introducing a Weibull model for shot duration in professional movie trailers. A similar approach was followed in [13] for home-video shot duration. While in the first case, shot duration is related to the creation of narrative atmospheres [10], [22], in the second one it constitutes an expression of human interest. However, this feature was not used in [13], as it was claimed that shot duration did not appear to be related to frame significance.

We argue that not only shots but also home-video clusters have clear temporal patterns. Unlike [13], we have made use of this information. Fig. 1(a) illustrates the empirical distribution of shot duration in the database and its approximation by a GMM. It can be seen that the duration of shots remains in the range of a couple of minutes. This is an indication of the typical amount of time that people are able to stay focused on when operating a camera. This limitation in interest is also evident by looking at the duration of video clusters. Fig. 1(b) shows the empirical distribution of cluster duration and its GMM approximation. Video clusters have a definite trend to last only a few minutes. In our database, approximately 95% of the clusters last less than 10 min. As a consequence of lack of attention, long video clusters are rare.

The complementary information is the distribution of number of shots per cluster. Although both the number of shots and the number of clusters per sequence vary considerably, most clusters are composed of only a few shots. Fig. 2(a) and (b) shows the distribution of shots and clusters per sequence. As a general trend, outdoor shots are shorter than indoor shots; hence,

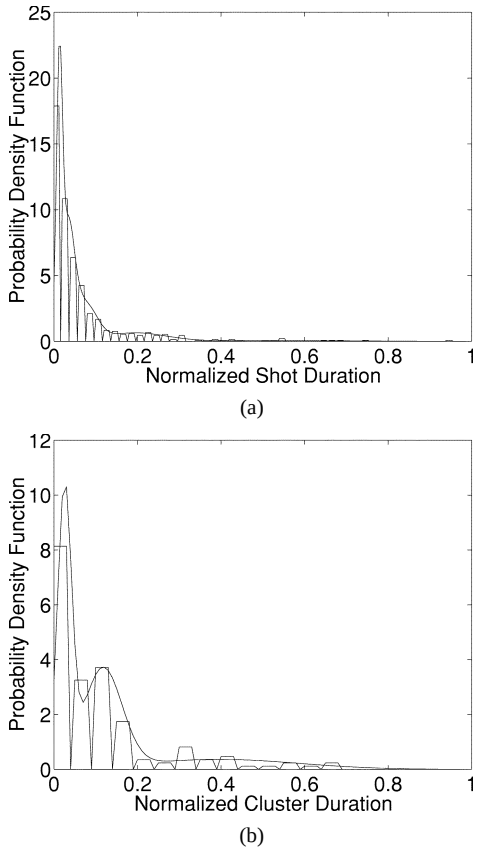

Fig. 1. (a) Consumer video shot duration. Empirical distribution of normalized shot duration and its GMM approximation. Shot duration was normalized by the longest shot in the database (580 s.). (b) Consumer video cluster duration. Distribution of normalized cluster duration and its GMM. The maximum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

outdoor 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. Fig. 2(c) shows the distribution of number of shots per cluster. 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.

2) Effect of Continuity: Home-video clusters composed of nonadjacent shots, that is clusters with forward and backward temporal jumps, are infrequent (about 3% of the clusters in the data set). In other words, clusters are localized in time, and therefore strong connectivity can be assumed for clustering, which has the benefits of computational simplicity.

3) Visual Similarity in Home-Video Clusters: Computing similarity between images/videos has been extensively addressed in CBIR [20]. An important question regards the visual structure of home-video clusters: how similar (respectively, dissimilar) are segments that belong to the same (respectively, a different) cluster? Let s_i and s_j denote two segments in a sequence, $\mathcal{E}$ denote a binary random value that indicates their belonging to the same cluster Ω ($\mathcal{E} = 1$ if $\Omega(s_i) = \Omega(s_j)$, and zero otherwise), and $d(s_i, s_j)$ denote a pairwise similarity measure. A B-bin mean RGB color histogram $h_i = \{h_{iz}\} \in \mathcal{R}^B$ was computed for each shot in the database, and used to construct the empirical distributions of intra-cluster $(p(d\,|\,\mathcal{E} = 1, \mathcal{I}))$ and inter-cluster

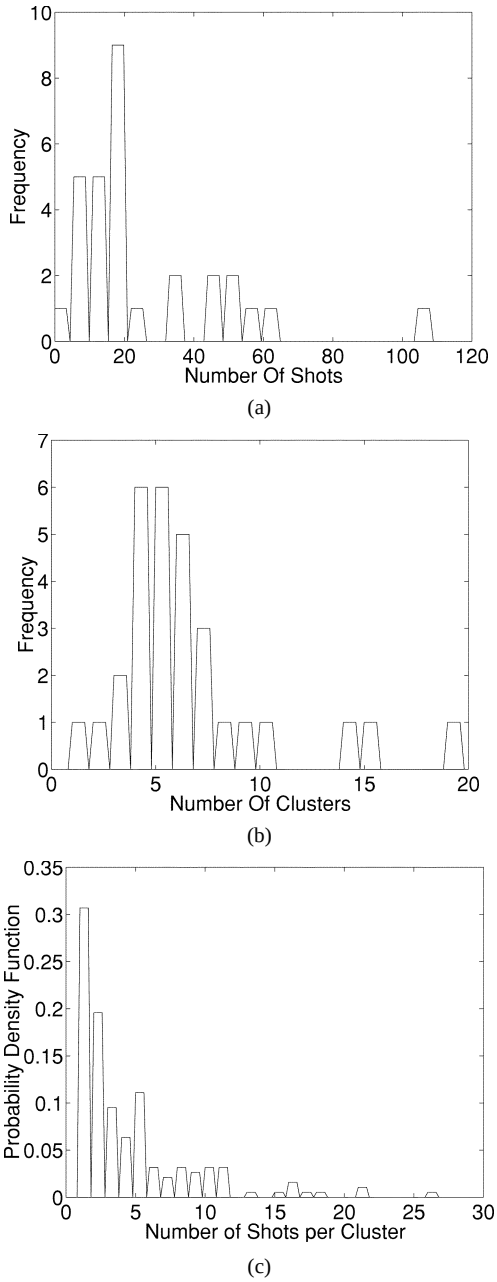

Fig. 2. Empirical distributions of: (a) number of shots per sequence; (b) number of clusters per sequence; and (c) distribution of shots per cluster. 50.3% of the clusters in the database are composed of one or two shots; 80.4% of the clusters are composed of six or less shots.

$(p(d\,|\,\mathcal{E} = 0, \mathcal{I}))$ pairwise visual similarity ($\mathcal{I}$ denotes the knowledge about the world). For the inter-cluster case, pairwise computation was limited within the interval that contains 95% of the probability mass of cluster duration [Fig. 1(b)]. The similarity measure was the typical norm in the L_1 space [20]. The distributions appeared quite overlapped as a result of the unrestricted content of home video [Fig. 3(a)]. This result highlights the limitations of both features and distance measures to define similarity among video segments and the challenges of the problem.

Recapitulating, the data analysis shows that there indeed exists cluster structure in home videos. It also suggests the development of methods that integrate segment visual similarity and

duration in a joint model, rely on strong temporal adjacency, and account for the fact that clusters are composed of few segments. One such method is described in the next section.

V. Our Approach

HAC algorithms can be based on probability models [2]. We propose to build models of visual similarity, duration, and temporal adjacency defined on pairs of segments. 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. The formulation as a two-class classification problem allows for the use of Bayesian decision theory. The MAP criterion establishes that given a realization x_{ij} of X (representing features extracted from segments s_i and s_j), the class $\mathcal{E}$ that must be selected is

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

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

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

where $p(x \,|\, \mathcal{E}, \mathcal{I})$ are the class-conditional pdfs of the observed features, $\Pr(\mathcal{E} \,|\, \mathcal{I})$ is the class prior, L denotes the posterior odds ratio, H_1 denotes the hypothesis that the segment pair belongs to the same cluster, and H_0 denotes the opposite. The prior allows for the introduction of knowledge about home video. The algorithm treats each elementary segment (shot) as a cluster, successively evaluates the pair of segments that corresponds to the largest L, merges when $L \geq 1$, and continues until H_1 in (1) is no longer valid. This greedy strategy bears similarity with the highest confidence first (HCF) method used in Bayesian image analysis [3]: at each step, decisions are made based on the piece of information that has the highest certainty. The formulation does not require any *ad-hoc* parameter determination, and can be seen as a generalization of previous time-constrained clustering algorithms [23]. 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, as described in [8].

The methodology requires the determination of an appropriate feature space, and the selection of models for the distributions. These issues are described in the following sections.

VI. Video-Segment Feature Extraction and Selection

First, shot boundaries are detected by standard methods [6]. Oversegmentation due to illumination or noise artifacts can be handled by the clustering algorithm. In the following, 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 home-video segments.

A. Extraction of Visual Features

Home-video shots usually contain more than one appearance, due to hand-held camera motion. We have adopted an approach that detects subshots inside each shot, which approximately correspond to individual scene appearances, and then extracts features from a set of random frames in each subshot. More elaborate key-frame extraction algorithms would not outperform random frame selection unless there was theoretical support for it. A shot s_i is defined as a collection of K subshots $s_i = \{s_{ik}, k \in \{1, \ldots, K\}\}$, and each subshot is characterized by a set of M random frames $s_{ik} = \{s_{ikm}, m \in \{1, \ldots, M\}\}$.

Subshots are sequentially extracted by thresholding a pairwise similarity measure between frames inside each shot [19]. Furthermore, subshots of very short duration are discarded, as they often correspond to fast camera pannings. For image representation, we have selected joint histograms of color and scene structure information [17]. Investigated features included:

1) color in RGB space (uniformly quantized to $8 \times 8 \times 8$ bins), and HSV space (vector-quantized to 1024 colors);
2) color ratios (known to be illumination-invariant), non linearly quantized to 32 levels [1];
3) edge density and edge direction features [21].

Regarding the similarity measure, if 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 subshots is defined as

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

where measures like the L_1 norm d_{L_1}, the Bhattacharyya coefficient metric d_{BT} [4], or the correlation coefficient measure d_{CC}, can be used for 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 subshots

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

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

$$d(s_i, s_j) = \min\{d(s_{ik}, s_{jl}), \quad \forall k \in \{1, \ldots, K\}, \quad l \in \{1, \ldots, L\}\}. \tag{2}$$

B. Selection of Visual Features

We estimated the intra- and inter-cluster distributions for all the features and similarity measures discussed in the previous subsection, using a subset of 75% of the sequences in our database. Features were selected based on the overlap they induced between the two pdfs. The empirical probability of error, for a noninformative prior, can be computed by $\Pr(e \,|\, \mathcal{I}) = (1/2)(\Pr(e \,|\, \mathcal{E}=0, \mathcal{I}) + \Pr(e \,|\, \mathcal{E}=1, \mathcal{I}))$, 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. Tables I and II summarize the results.

Table I shows the empirical probability of error computed for RGB histograms with and without subshot detection, and for four-dimensional (4-D) histograms that combine color and edge density (EDEN), edge directions (EDIR), and color ratios (YR). The advantage of using subshot detection and random frames

TABLE I
FEATURE-SELECTION: L_1 METRIC

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 II
COMPARISON OF SIMILARITY MEASURES. JOINT HISTOGRAM RGB-EDEN

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

(SS+RF) as opposed to global shot information is evident, as subshot analysis has improved the separation between the two classes. The use of joint histograms further improves discrimination. RGB-EDEN produced slightly better results than the other 4-D histograms. No improvement was found when using the HSV color model. Additionally, the results of applying various similarity measures are presented in Table II and Fig. 3(b). The L_1 norm and the metric based on Bhattacharyya coefficient produced better results than the correlation coefficient. The Bhattacharyya coefficient can be interpreted as the cosine of the angle between the component-wise square-rooted pdfs approximated from the joint histograms [4], so d_{L_1} and d_{BT} can be seen as representations of magnitude and angle, and constitute the features to characterize visual similarity.

C. Selection of Temporal Features

The analysis in Section IV made evident the possibility of using strong adjacency for clustering. The accumulated duration of two segments is an indication of their belonging to the same cluster (segments of increasing length become less likely to belong to the same cluster). Such a feature is defined by

$$\Delta_{ij} = \min\{|e_j - b_i|, |e_i - b_j|\} \tag{3}$$

where b_i and e_i denote the first and last frame of s_i [19].

VII. MODELING OF LIKELIHOOD FUNCTIONS AND PRIOR

A. Modeling of Likelihood Functions With GMMs

The features described define a space, with vectors $X = (d_{L_1}, d_{\mathrm{BT}}, \Delta) \in \mathcal{X}$. The class-conditional pdfs of the observed features are represented by multivariate GMMs

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

where $N_{\mathcal{E}}$ is the number of components in each mixture, ω_i denotes the prior of the i-th component, $p(x \,|\, \mathcal{E}, \theta_i, \mathcal{I}) = \mathcal{N}(\mu_i, \Sigma_i)$ is a three-dimensional (3-D) Gaussian with full covariance matrix, parameterized by $\theta_i = \{\mu_i, \Sigma_i\}$, and $\Theta = \{\{\omega_i\}, \{\theta_i\}\}$ denotes all the parameters. Expectation-maximization (EM) constitutes the standard procedure for maximum likelihood (ML) estimation for GMMs [5]. Additionally, model selection is performed via minimum description length (MDL) by choosing

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

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

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

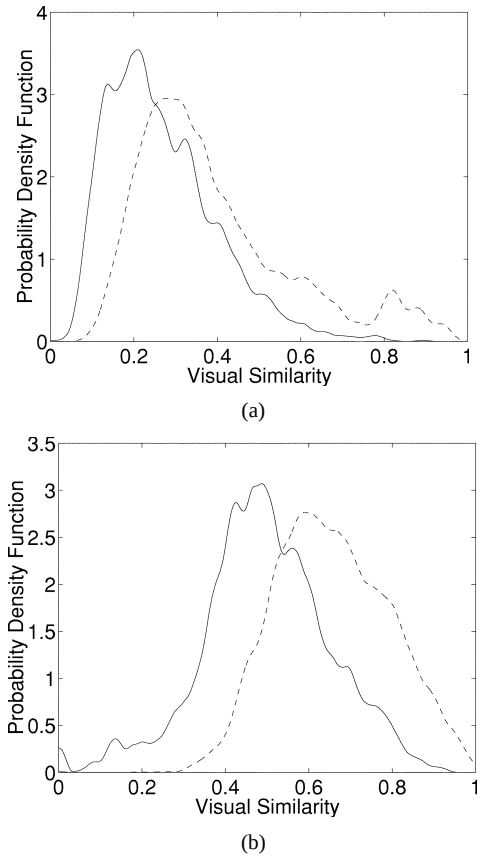

Fig. 3. Pairwise shot visual similarity distributions. Intra-cluster and inter-cluster pdfs are represented by continuous and dotted-line curves, respectively. (a) Features based on global shot information (RGB mean histograms); L_1 norm. (b) Features from subshot detection, random frame extraction, and joint RGB-EDEN histograms; similarity measure based on Bhattacharyya coefficient.

B. Modeling of Prior

The prior encodes the belief about the clustering process [9]. While the simplest assumption is a uniform prior, Section IV suggested that merging must be discouraged as clusters usually consist of only a few shots. The prior should reflect this knowledge. One possibility is to determine it from the available evidence. While this technique does not conform to the Bayesian principle, it usually produces better solutions than arbitrary priors. Assuming independence among the N training data, the ML estimator of the prior is defined by

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

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

VIII. 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. Although many algorithms for video segment clustering have been proposed in the literature [25], [23], [19], [14], their performance using the two mentioned criteria is unknown in several cases. Refer to [8] for a review of the literature in this respect.

A. Ground Truth

A third-party cluster ground truth was determined based on human evaluation of shot visual similarity, temporal adjacency, and blind context understanding. This type of ground truth is common for performance evaluation, including movie analysis and still images [15], and is useful to perform benchmarking against the limit of a computer algorithm which has no context knowledge. Note that although there are differences of judgement between people due to the uncertainty about the contents, there indeed exists cluster structure in home videos. The incorporation of multiple human judgements of similarity and the use of statistical measures for evaluation of video analysis algorithms are research issues currently under study.

B. Performance-Evaluation Procedure

Results were generated with the leave-one-out method: one video sequence was held for evaluation while the rest were included in the training set. Given NC, the number of clusters in the ground truth (either for an individual sequence or for 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 $\mathrm{FP} = \mathrm{DC} - \mathrm{NC}$ if $\mathrm{DC} - \mathrm{NC} \geq 0$ and zero otherwise, and *false negatives* (FN), defined by $\mathrm{FN} = \mathrm{NC} - \mathrm{DC}$ if $\mathrm{DC} - \mathrm{NC} \leq 0$ 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. The performance figures are then turned into probabilities. If z is any of the parameters of interest, the frequentist performance evaluation 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. The first measure gives the same importance to each shot (or cluster) in the database; the second one 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 $\mathrm{FP} = 0$, the figures are computed by

$$\mathrm{dc} = \frac{\mathrm{DC}}{\mathrm{NC}}; \quad \mathrm{fp} = 0; \quad \mathrm{fn} = \frac{\mathrm{FN}}{\mathrm{NC}} \tag{4}$$

and if $\mathrm{FN} = 0$, the expressions are

$$\mathrm{dc} = \frac{\mathrm{DC} - \mathrm{FP}}{\mathrm{DC}}; \quad \mathrm{fp} = \frac{\mathrm{FP}}{\mathrm{DC}}; \quad \mathrm{fn} = 0. \tag{5}$$

Additionally

$$\mathrm{sie} = \frac{\mathrm{SIE}}{\mathrm{NS}}; \quad \mathrm{co} = \frac{\mathrm{CO}}{\mathrm{NS}} \tag{6}$$

where NS stands for the number of shots. For micro-averages, (4)–(6) are valid for each individual sequence. Results are then accumulated and averaged over the whole database.

C. Results

The detailed results are shown in Table III (videos sorted according to number of shots). The summarized results appear in Tables IV and V. Table IV shows the capability of our methodology to detect clusters. This is a hard problem, due to the variability in the data set [Fig. 2(b)]. The macro-average shows that the total number of detected clusters approximately corresponds to the 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. In contrast, the micro-average is a more reliable 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 reflects the fact that the method has a tendency to oversegment (from Table III, 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 [23], [19]. Furthermore, several of the false negatives actually consist of only one or two shots according to the ground truth. We also show the poor result obtained with an algorithm that randomly estimates the number of clusters for each video. This result simulates the case in which home videos truly did not have structure, so any clustering would be equally good.

Table V describes the performance 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 measures are useful. Variations between them indicate difference of performance from sequence to sequence. We selected a number of baseline methods for comparison, which assume the *correct* number of clusters for each sequence, as dictated by the ground truth. The methods are: 1) B_1, which assigns a uniform and temporally adjacent number of shots per cluster [18]; 2) B_2, a version of K-means for shots, in which the centroids were initialized with randomly selected shots from each sequence; and 3) B_3, the same variation of K-means, but in which the centroids were initialized with equally "spaced" shots (in terms of shot number). The "distance" between a shot and a centroid in the

TABLE III
VIDEO-CLUSTERING RESULTS ON KODAK HOME-VIDEO DATABASE

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

TABLE IV
CLUSTER DETECTION PERFORMANCE

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

TABLE V
SHOT ASSIGNMENT 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

K-means algorithm was computed by (2). 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 considered the case of random clustering (B_0).

The results show that our methodology outperformed all of the baseline methods. Using macro-averages (respectively, micro-averages) as measurement, our methodology assigned 71.1% (respectively, 71.4%) $[100 * (1 - \text{sie})]$ of the shots to the correct cluster. In contrast, the two K-means algorithms produced a similar performance, the best one generating 47.6% (respectively, 56%) of correct shot assigments. Interestingly, uniform shot-assigment performed better than K-means [54.7% (respectively, 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 3.55 (respectively, 4.62) operations are needed in average to correct the cluster assignments in a 20-min video with the proposed method.

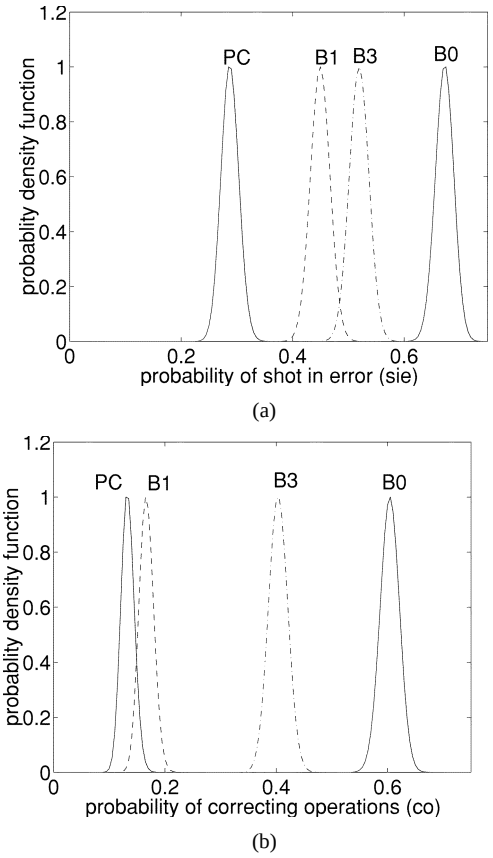

Fig. 4. (a) Posterior distributions of the probability of shot in error, for uniform prior and different structuring algorithms. PC denotes our approach. (b) Posteriors of the probability of correcting operations.

TABLE VI
EFFECT OF PRIOR PROBABILITY

	dc_M	fp_M	fn_M	sie_M	dc_m	fp_m	fn_m	sie_m
U	0.470	0.000	0.529	0.393	0.573	0.000	0.427	0.309
E	0.934	0.065	0.000	0.289	0.750	0.171	0.079	0.286

The Bayesian approach can be used 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 [9]. In the N-shot database, suppose n shots in error are observed. The likelihood $\Pr(n \mid \text{sie})$ is a binomial distribution, $\Pr(n \mid \text{sie}) \propto \text{sie}^n(1 - \text{sie})^{N-n}$. If, for analytic convenience, it is further assumed a uniform prior $p(\text{sie})$, the expression for the posterior becomes

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

Fig. 4(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. 4(b) presents the corresponding analysis to compare the posterior distributions of the probability of correcting operations $p(\text{co} \mid n)$.

The effect of the prior distribution in the clustering algorithm is shown in Table VI. A uniform prior (U) does not make use

Fig. 5. Generated structure for a *Baby* video sequence (detail). Each shot is represented by one frame. Clusters correspond to rows of shots.

of knowledge of the problem: merging should be discouraged as most video clusters consist of a few 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 $\mathrm{dc} = 0.573$, 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, with most errors coming from shots that belong to small clusters which were erroneously merged. On the other hand, the ML estimate of the prior (E) was $\Pr(\mathcal{E} \mid \mathcal{I}) = \{0.87, 0.13\}$. This distribution reflected the knowledge about the problem in better terms, and improved performance.

One example of the generated clusters is shown in Fig. 5. Each cluster is displayed as a row of shots, which are in turn represented by a random frame each. Qualitatively, the methodology provides quite reasonable results. We have integrated it in the development of a system for organization of home videos. An interface that displays the structure as a tree consisting of sequence, cluster, shot, and subshot levels is shown in Fig. 6. The interface also allows for reorganization of video structures, and retrieval of information from them.

D. Limitations

There are three main reasons for erroneous merging: high visual similarity between semantically disjoint but temporally adjacent clusters, shots of very short duration, and clusters of very short duration. Furthermore, the two reasons for erroneous oversegmentation are high intra-cluster visual variability, and unusu-

Fig. 6. A video structure as a tree. The root node corresponds to the sequence, the middle nodes to the clusters, and the leaf columns represent the shots, composed of random frames extracted from subshots.

ally long clusters (see [8] for examples). As a general trend, outdoor scenes are harder to cluster correctly.

Although the proposed approach has produced good results, it is known that the use of global low-level features has limitations to model semantic information [20]. Our work could benefit from the use of image segmentation into a few regions as the starting point for matching elements across representative frames. One advantage of the proposed method 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, the introduction of higher-level features such as faces should also be investigated [16].

IX. Concluding Remarks

This paper presented a methodology to discover cluster structure in home videos by incorporating some of 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. Features of intra- and inter-cluster visual similarity, adjacency, and duration were exploited. The obtained results are encouraging, but also illustrate the complexity of the research problem. Several issues remain open, including the investigation of both better mechanisms to quantify similarity between video segments and features that can capture such similarity, and the integration of region-based and multimedia representations (i.e., using audio) in the proposed framework.

Acknowledgment

The authors thank P. Stubler for providing software for shot-boundary detection, S. Ruiz-Correa for discussions, N. Triroj for help with data collection, the Eastman Kodak Company for the database, and the anonymous reviewers for their comments.

References

[1] D. A. Adjeroh and M. C. Lee, "On ratio-based color indexing," *IEEE Trans. Image Processing*, vol. 10, pp. 36–48, Jan. 2001.
[2] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, no. 3, pp. 803–821, Sept. 1993.
[3] P. Chou and C. Brown, "The theory and practice of Bayesian image labeling," *Int. J. Comput. Vis.*, vol. 4, pp. 185–210, 1990.
[4] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of nonrigid objects using mean shift," *Proc. IEEE CVPR*, June 2000.
[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, ser. B, vol. 39, pp. 1–38, 1977.
[6] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 1–13, Feb. 2000.

[7] D. Gatica-Perez, M.-T. Sun, and A. Loui, "Consumer video structuring by probabilistic merging of video segments," *Proc. IEEE ICME*, Aug. 2001.
[8] ——, "Finding Structure in Home Videos by Probabilistic Hierarchical Clustering,", IDIAP Tech. Rep. RR-02-22, May 2002.
[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. London, U.K.: Chapman & Hall, 1996.
[10] J. Hart, *Film Directing Shot by Shot: Visualizing from Concept to Screen*. Studio City, CA: Michael Wiese Productions, 1991.
[11] G. Iyengar and A. Lippman, "Content-based browsing and edition of unstructured video," *Proc. IEEE ICME*, Aug. 2000.
[12] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. New York: Prentice-Hall, 1998.
[13] J. R. Kender and B. L. Yeo, "On the structure and analysis of home videos," in *Proc. Asian Conf. Computer Vision*, Taipei, Taiwan, R.O.C., Jan. 2000.
[14] R. Lienhart, "Abstracting home video automatically," in *Proc. ACM Multimedia Conf.*, Orlando, FL, Oct. 1999, pp. 37–41.
[15] A. Loui and M. Wood, "A software system for automatic albuming of consumer pictures," in *Proc. ACM Multimedia 99*, Orlando, FL, Nov. 1999.
[16] W.-Y. Ma and H. J. Zhang, "An indexing and browsing system for home video," in *Proc. EUSIPCO*, Patras, Greece, 2000, pp. 131–134.
[17] G. Pass and R. Zabih, "Comparing images using joint histograms," *ACM J. Multimedia Syst.*, vol. 7, no. 3, pp. 234–240, May 1999.
[18] J. Platt, "AutoAlbum: Clustering digital photographs using probablisitic model merging," *Proc. IEEE Workshop on Content-Based Access to Image and Video Libraries*, 2000.
[19] Y. Rui and T. Huang, "A unified framework for video browsing and retrieval," in *Image and Video Processing Handbook*, A. Bovik, Ed. New York: Academic, 2000, pp. 705–715.
[20] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1349–1380, Dec. 2000.
[21] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang, "Image classification for content-based indexing," *IEEE Trans. Image Processing*, vol. 10, pp. 117–130, Jan. 2001.
[22] N. Vasconcelos and A. Lippman, "A bayesian video modeling framework for shot segmentation and content characterization," *Proc. IEEE Computer Vision and Pattern Recognition*, 1997.
[23] M. Yeung, B. L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Comput. Vis. Image Understand.*, vol. 71, no. 1, pp. 94–109, July 1998.
[24] 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*, vol. 4315, Jan. 2001, pp. 262–269.
[25] 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*, vol. 2670, Feb. 1996, pp. 239–246.