

Consumer Video Structuring by Probabilistic Merging of Video Segments

Daniel Gatica-Perez, Ming-Ting Sun
Department of Electrical Engineering
University of Washington
Box 352142, Seattle, WA 98195

Alexander Loui
Research and Development Laboratories
Eastman Kodak Company
Rochester, NY 14650-1816

ABSTRACT

Accessing, organizing, and manipulating home videos constitutes a technical challenge due to their unrestricted content and the lack of storyline. In this paper, we present a methodology for structuring consumer video, based on the development of statistical models of similarity and adjacency between video segments in a probabilistic formulation. Learned Gaussian mixture models of inter-segment visual similarity, temporal adjacency, and segment duration are used to represent the class-conditional densities of observed features. Such models are then used in a sequential merging algorithm consisting of a binary Bayes classifier, where the merging order is determined by a variation of Highest Confidence First (HCF), and the merging criterion is Maximum a Posteriori (MAP). The merging algorithm can be efficiently implemented and does not need any empirical parameter determination. Finally, the representation of the merging sequence by a tree provides for hierarchical, nonlinear access to the video content. Results on an eight-hour home video database illustrate the validity of our approach.

1. Introduction

Among all sources of video content, consumer video probably constitutes the one that most people are or would eventually be interested in dealing with. Efficient tools for accessing, organizing, and manipulating the huge amount of raw information contained in personal video materials open doors to the organization of video events in albums, video baby books, edition of postcards with stills extracted from video data, multimedia family web pages, etc. [7-9]. In fact, the variety of user interests asks for an interactive solution, which requires a minimum amount of user feedback to specify the desired tasks at the semantic level, and that provides automated algorithms for those tasks that are tedious or can be performed reliably.

Unrestricted content and the absence of *storyline* are the main characteristics of home video. Consumer contents are usually composed of a set of events, either isolated or related, each composed of one or a few shots, randomly spread along time. Such characteristics make consumer video unsuitable for video analysis approaches based on storyline models [4]. However, there still exists a *spatio-temporal* structure, based on visual similarity and temporal adjacency between video segments (sets of shots) that appears evident after a statistical analysis of a large home video database. Such structure, essentially equivalent to the structure of consumer still images [9], points towards addressing home video structuring as a problem of clustering. This has indeed been the direction taken by most research in video analysis, even when dealing with storylined content. Using shots as the fundamental unit of video structure, K-means [15] distribution-based clustering [7], and time-constrained merging techniques [14], [11] have been tested. As a byproduct, clustering allows for the generation of hierarchical

representations for video content, which provide nonlinear access for browsing and manipulation.

In this paper, we investigate statistical models of visual and temporal features in consumer video for organization purposes. A Bayesian formulation seems appealing to encode prior knowledge of the *spatio-temporal* structure of home video. We propose a methodology that uses video shots as the unit of organization and that supports the creation of a video hierarchy for interaction. Our approach is based on an efficient probabilistic video segment merging algorithm which integrates inter-segment features of visual similarity, temporal adjacency, and duration in a joint model, that allows for the generation of video clusters without empirical parameter determination.

To date, only a few works have dealt with analysis of home video [7], [8], [11]. With the exception of [7], none of the previous approaches have analyzed in detail the inherent statistics of such content. From this point of view, our work is more related to the work in [13], that proposes a Bayesian formulation for shot boundary detection, and to the work in [7], that addresses home video analysis with a different formulation.

The rest of the paper is organized as follows. Section 2 describes our approach in details. Section 3 presents results on a home video database. Section 4 draws some concluding remarks.

2. Overview of our Approach

Assume a feature vector representation for video segments, i.e., suppose that a video clip has been divided into shots or segments, and that features that represent them have been extracted. Any clustering procedure should specify mechanisms both to assign cluster labels to each segment in the home video clip and to determine the number of clusters. The clustering process needs to include time as a constraint, as video events are of limited duration [14], [11]. However, the definition of a generic generative model for *intra-segment* features in home videos is particularly difficult, given their unconstrained content. Instead, we propose to analyze home video using statistical *inter-segment* models. In other words, we propose to build up models that describe the properties of visual and temporal features defined on *pairs of segments*. Inter-segment features naturally emerge in a *merging* framework, and integrate visual dissimilarity, duration, and temporal adjacency. A merging algorithm can be thought of as a classifier, which sequentially takes a pair of video segments and decides whether they should be merged or not. Let s_i and s_j denote the i -th and j -th video segments in a video clip, and let ε be a binary r.v. that indicates whether such pair of segments correspond to the same cluster and should be merged or not. The formulation of the merging process as a sequential two-class (merge/not merge) pattern classification problem allows for the application of concepts from Bayesian decision theory [3]. The Maximum a Posteriori (MAP) criterion

establishes that given an n-dimensional realization x_{ij} of an r.v. x (representing inter-segment features and detailed later in the paper), the class that must be selected is the one that maximizes the *a posteriori* probability mass function of ε given x , i.e.,

$$\varepsilon^* = \arg \max_{\varepsilon} \Pr(\varepsilon | x)$$

Applying Bayes rule, the MAP principle can be expressed as

$$p(x | \varepsilon = 1) \Pr(\varepsilon = 1) \underset{H_0}{\overset{H_1}{>}} p(x | \varepsilon = 0) \Pr(\varepsilon = 0)$$

where $p(x | \varepsilon)$ is the class-conditional pdf (likelihood) of x given ε , $\Pr(\varepsilon)$ is the prior of ε , H_1 denotes the hypothesis that the pair of segments should be merged, and H_0 denotes the opposite. With this formulation, the classification of pairs of shots is performed sequentially, until a certain stop criteria is satisfied. Therefore, the tasks are the determination of a useful feature space, the selection of models for the distributions, and the specification of the merging algorithm. Each of these steps are described in the following.

2.1. Video Segmentation

To generate the basic segments, shot boundary detection is computed by a series of methods to detect the cuts usually found in home video [5]. Over-segmentation due to detection errors (e.g. due to illumination or noise artifacts) can be handled by the clustering algorithm. Additionally, videos of very poor quality are removed.

2.2. Video Inter-segment Feature Definition

Both visual dissimilarity and temporal information have been used for clustering in the past [14], [11]. In the first place, in terms of discerning power of a visual feature, it is clear that a single frame is often insufficient to represent the content of a segment. From the several available solutions, we have selected the *mean segment color histogram* to represent segment appearance. The L1 norm of the mean segment histogram difference is used to visually compare segments,

$$\alpha_{ij} = \sum_{k=1}^B |m_{ik} - m_{jk}|$$

where α_{ij} denotes visual dissimilarity between segments, B is the number of histogram bins, and m_{ik} is the value of the k -th bin of the mean color histogram of segment s_i .

In the second place, the *temporal separation* between segments s_i and s_j is defined as

$$\beta_{ij} = \min(|e_i - b_j|, |e_j - b_i|)(1 - \delta_{ij})$$

where δ_{ij} denotes a Kronecker's delta, and b_i, e_i denote first and last frame of segment s_i . Additionally, the combined duration of two individual segments is also a strong indication about their belonging to the same cluster. Fig. 1 shows the empirical distribution of home video shot duration for approximately 660 shots from our database with ground-truth, and its fitting by a Gaussian mixture model (see next subsection). Even though videos correspond to different scenarios and were

filmed by multiple people, a clear temporal pattern is present [13]. The *accumulated segment duration* τ_{ij} is defined as

$$\tau_{ij} = \text{card}(s_i) + \text{card}(s_j)$$

where $\text{card}(s)$ denotes the number of frames in segment s .

Our method provides a probabilistic alternative for previous techniques that relied on similar features [14], [11] for clustering. Other features can be easily integrated in the formulation.

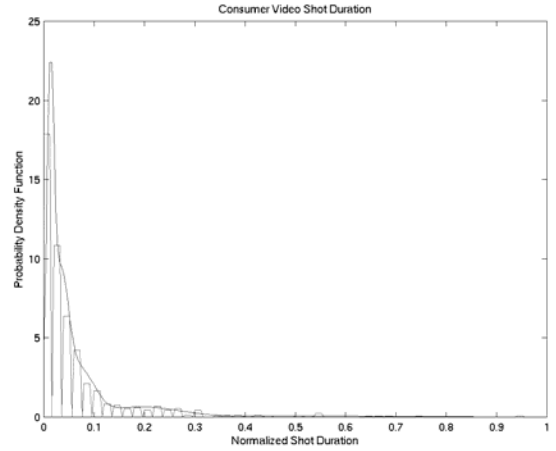


Fig 1. Modeling home video shot duration. The empirical distribution, and an estimated Gaussian mixture model consisting of six components, are superimposed. Duration was normalized to the longest duration found in the database (580 sec).

2.3. Modeling of Likelihoods and Priors

The described features become the components of the feature space X , with vectors $x = (\alpha, \beta, \tau)$. To analyze the separability of the two classes, Fig. 2 shows a scattering plot of 4000 labeled inter-segment feature vectors extracted from home video. Half of the samples correspond to hypothesis H_1 (light gray), and the other half to H_0 (dark gray). The features have been normalized.

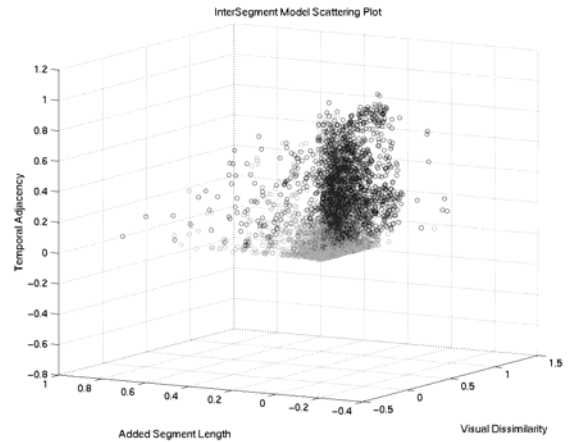


Fig 2. Scattering plot for training inter-segment feature vectors.

The plot indicates that the two classes are in general separated. A projection of this plot clearly illustrates the limits of relying on pure visual similarity. We have adopted a parametric mixture model for each of the class-conditional densities of the observed inter-segment features,

$$p(x | \varepsilon, \Theta) = \sum_{i=1}^{K_\varepsilon} \Pr(c = i) p(x | \varepsilon, \theta_i)$$

where K_ε is the number of components in each mixture, $\Pr(c = i)$ denotes the prior probability of the i -th component, $p(x | \varepsilon, \theta_i)$ is the i -th pdf parameterized by θ_i , and $\Theta = \{\Pr(c), \{\theta_i\}\}$ represents the set of all parameters. In this paper, we assume multivariate Gaussian forms for the components of the mixtures in d -dimensions

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

so that the parameters θ_i are the means μ_i and covariance matrices Σ_i [3]. The expectation-maximization (EM) algorithm constitutes the standard procedure for Maximum Likelihood estimation (ML) of the set of parameters Θ [2]. EM is a technique for finding ML estimates for a broad range of problems where the observed data is in some sense incomplete. In the case of a Gaussian Mixture, the incomplete data are the unobserved mixture components, whose prior probabilities are the parameters $\{\Pr(c)\}$. EM is based on increasing the conditional expectation of the log-likelihood of the complete data given the observed data by using an iterative hill-climbing procedure. Additionally, model selection, i.e., the number of components of each mixture can be automatically estimated using the Minimum Description Length (MDL) principle [10].

Instead of imposing independence assumptions among the variables, the full joint class-conditional pdfs are estimated. The ML estimation of the parametric models for $p(x | \varepsilon = 0)$ and $p(x | \varepsilon = 1)$, by the procedure just described, produces probability densities represented by ten components in both cases, respectively.

Finally, the discrete prior probability mass functions $\Pr(\varepsilon)$, which encode the knowledge or belief about the merging process characteristics (home video clusters mostly consist of only a few shots), are ML-estimated from the available training data [3].

2.4. Video Segment Clustering

Any merging algorithm requires three elements: a feature model, a merging order, and a merging criterion [6]. In our proposed algorithm, the class-conditionals are used to define both the merging order and the merging criterion. Merging algorithms can be efficiently implemented by the use of adjacency graphs and hierarchical queues, which allow for prioritized processing. Their use in Bayesian image analysis first appeared in [1] with the Highest Confidence First (HCF) optimization method. The concept is intuitively appealing: at each step, decisions should be made based on the piece of information that has the highest certainty. Recently, similar formulations have appeared in [6] in morphological processing.

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

Queue initialization. At the beginning of the process, inter-shot features x_{ij} are computed. Each feature x_{ij} is introduced in the queue with priority equal to the probability of merging the corresponding pair of shots, $\Pr(\varepsilon = 1 | x_{ij})$.

Queue depletion/updating. Our definition of priority allows making decisions always on the pair of segments of highest certainty. Until the queue is empty, the procedure is as follows:

1. Extract an element from the queue. This element (pair of segments) is the one that has the highest priority.
2. Apply the MAP criterion to merge the pair of segments, i.e.,

$$p(x_{ij} | \varepsilon = 1) \Pr(\varepsilon = 1) > p(x_{ij} | \varepsilon = 0) \Pr(\varepsilon = 0)$$

3. 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 1. Otherwise (H_0), go to step 1.

When a pair of segments is merged, the model of the new segment s_i is updated by

$$m_i = (\text{card}(s_i)m_i + \text{card}(s_j)m_j) / (\text{card}(s_i) + \text{card}(s_j))$$

$$b_i = \min(b_i, b_j)$$

$$e_i = \max(e_i, e_j)$$

$$\text{card}(s_i) = \text{card}(s_i) + \text{card}(s_j)$$

After having updated the model of the (new) merged segment, four functions need to be implemented to update the queue:

1. Extraction from the queue of all those elements that involved the originally individual (now merged) segments.
2. Computation of new inter-segment features $x = (\alpha, \beta, \tau)$ using the updated model.
3. Computation of new priorities $\Pr(\varepsilon = 1 | x_{ij})$.
4. Insertion in the queue of elements according to new priorities.

Note that, unlike previous methods, our formulation does not need any empirical parameter determination [14],[11].

The merging sequence, i.e., a list with the successive merging of pairs of video segments, is stored and used to generate a hierarchy. Furthermore, for visualization and manipulation, after emptying the hierarchical queue in the merging algorithm, further merging of video segments is allowed to build a complete merging sequence that converges into a single segment (the whole video clip). The merging sequence is then represented by a partition tree, which has proven to be an efficient structure for hierarchical representation of visual content [12], and provides the starting point for user interaction.

2.5. Video Hierarchy Visualization.

We have built a prototype of an interface to display the tree representation of the analyzed home video, based on key frames. A set of functionalities that allow for manipulation (correction, augmentation, reorganization) of the automatically generated video clusters, along with cluster playback, and other VCR capabilities is under development. An example of the tree representation appears in Fig. 3.

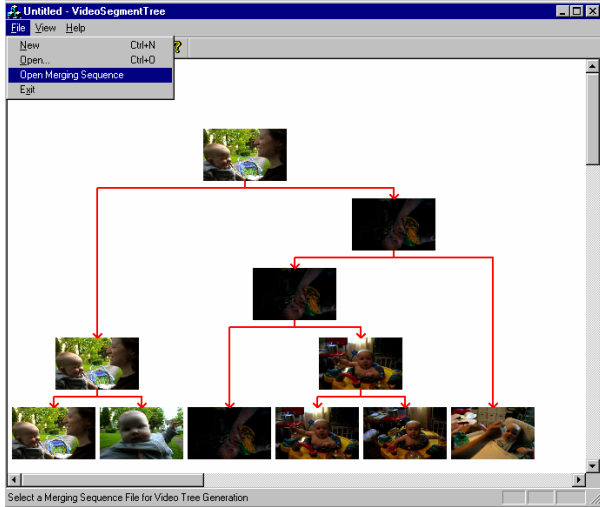


Fig 3. Displaying the Video Segment Tree.

3. Results

Our methodology was evaluated on a database of 24 home MPEG video clips of different characteristics. Each video clip has an approximate duration of 18-25 minutes. The total number of video shots in the database is 659, and the total duration is about 8 hours. A third party ground-truth at shot and cluster level was generated by hand. We used 20 video clips for training, while the rest were used for testing.

Table I shows the results for the testing set. *Detected Clusters (DC)* is self-explanatory. *False Positives (FP)* denotes the number of clusters detected by the algorithm but not included in the ground-truth, and *False Negatives (FN)* indicates the opposite. These are the figures traditionally reported in clustering experiments. However, to perform a more strict evaluation, we have included two more figures. *Shots In Error (SIE)* denotes the number of shots whose cluster label do not match the label in the ground-truth. Finally, *Correcting Operations (CO)* indicates the number of operations (merging, splitting, etc.) that are necessary to correct the results so that SIE is zero. We believe this is a good figure of the effort required in interactive systems.

Video-clip	Duration	Shots	DC	FP	FN	SIE	CO
Bubbles	19:56	12	4	0	0	1	1
Cindy	21:39	18	6	0	0	5	2
Clem	20:01	35	5	0	1	7	4
Sue	20:02	10	5	0	2	2	2
OT	20:10	18.8	5	0	0.8	3.8	2.3
OD	20:43	27.5	5.1	0.4	2.6	9.8	5.1

Table I. Home Video Clustering Results.

We see that for the testing clips, the merging process achieved reasonably good results. The analysis of the database shows that about 50% of the clusters consist of one or two shots. This fact and the large variety of content make home video especially hard to cluster. The overall results on the testing set (*OT*), and on the whole database (*OD*) are indicated in Table I. On average, only five operations are needed to correct the cluster assignments in a 20-minute home video. Most merging errors are

due to visually similar shots that are temporally adjacent but semantically disjoint. We believe this result is of good quality, especially because even human performance is uncertain when clustering consumer visual contents. In order to compensate for variability of human evaluation, one could define a probability distribution of human judgment, and evaluate the performance of automatic algorithms based, for instance, on posterior intervals.

4. Concluding Remarks

The obtained results show the validity of our approach. We are currently experimenting with other features of visual dissimilarity with better discrimination power. Additionally, we are studying incremental learning schemes to improve the classifier as more samples become available. Finally, performance evaluation of systems for access and organization of consumer content still constitutes an open issue. We are not aware either of any public consumer video database or of any comparative study of home video segment clustering techniques.

Acknowledgments

The authors thank Peter Stubler for providing software for shot boundary detection, and for valuable discussions.

References

- [1] P. Chou and C. Brown. "The Theory and Practice of Bayesian Image Labeling". *IJCV*, 4, pp. 185-210, 1990.
- [2] 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.
- [3] R.O.Duda, P.E. Hart, D. G. Stork. *Pattern Classification*. Second Edition. John Wiley and Sons, 2000.
- [4] S. Eickeler and S. Muller. "Content-based Video Indexing of TV Broadcast News Using HMMs". *Proc. ICASSP 99*, Phoenix, pp. 2997-3000.
- [5] U. Gargi, R.Kasturi, and S.H. Strayer. "Performance Characterization of Video-Shot-Change Detection Methods". *IEEE CSVT*, Vol. 10, No. 1, February 2000, pp. 1-13.
- [6] L. Garrido, P. Salembier, D. Garcia, "Extensive operators in partition lattices for image sequence analysis". *Sign. Proc.*, 66(2):157-180, 1998.
- [7] G. Iyengar, and A. Lippman. "Content-based browsing and edition of unstructured video". *IEEE ICME*, New York City, Aug. 2000.
- [8] R. Lienhart. "Abstracting Home Video Automatically". *ACM Multimedia Conference*, Orlando, Oct. 1999. pp. 37-41.
- [9] A. Loui and A. Savakis, "Automatic image event segmentation and quality screening for albuming applications," *IEEE ICME*, New York City, Aug. 2000.
- [10] J. Rissanen. "Modeling by shortest data description". *Automatica*, 14: 465-471, 1978.
- [11] Y. Rui and T.S. Huang. "A Unified Framework for Video Browsing and Retrieval". In A.C. Bovik, Ed. *Handbook of Image and Video Processing*. Academic Press, 1999.
- [12] 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.
- [13] N. Vasconcelos and A. Lippman. "A Bayesian Video Modeling Framework for Shot Segmentation and Content Characterization". *Proc. CVPR 1997*.
- [14] 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.
- [15] D. Zhong and H. J. Zhang. "Clustering Methods for Video Browsing and Annotation", in *Proc. IS&T/SPIE Storage and Retrieval for Still Images and Video Databases IV*, Feb. 1996, vol. 2670, pp.239-246.