IDIAP RESEARCH REPORT

Probabilistic Home Video Structuring: Feature Selection and Performance Evaluation

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

IDIAP–RR 02-11

April 17, 2002

to appear in
IEEE International Conference on Image Processing 2002

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
[b] Dept. of Electrical Engineering, University of Washington, Seattle, WA, USA
[c] Imaging Research Labs, Eastman Kodak Company, Rochester, NY USA

IDIAP Research Report 02-11

Probabilistic Home Video Structuring: Feature Selection and Performance Evaluation

Daniel Gatica-Perez Ming-Ting Sun Alexander Loui

April 17, 2002

to appear in
IEEE International Conference on Image Processing 2002

Abstract. We recently proposed a method to find cluster structure in home videos based on statistical models of visual and temporal features of video segments and sequential binary Bayesian classification. In this paper, we present analysis and improved results on two key issues: feature selection and performance evaluation, using a ten-hour database (30 video clips, 1,075,000 frames). From multiple features and similarity measures, visual features are selected in order to minimize the empirical probability of misclassification. Temporal features are chosen to reflect the patterns existing in both shot and cluster duration and adjacency. Finally, we describe a detailed performance evaluation procedure that includes cluster detection, individual shot-cluster labeling, and prior selection.

1 Introduction

The interest in developing efficient schemes for accessing and retrieving home video has increased [7], [5], [6], [8], [10], [4], in view of the amount of available information and the variety of applications. Home videos are composed of a set of *events*, each composed of a few video shots, randomly recorded along time. However, and in spite of this lack of storyline, recent studies of home video databases reveal that non-professional filmmakers implicitly follow certain rules of *attention focusing* and *recording*, which induce structure in the video content [6], [4].

Specifically, we have argued that the *cluster structure* of home videos can be disclosed from such rules, using a methodology based on two concepts: the development of statistical models of visual and temporal features of video segments, and the reformulation of hierarchical clustering as sequential binary Bayesian classification [4]. Such approach allows for the integration of prior knowledge of the cluster structure of home videos, and offers the advantages of a principled methodology [3]. We have shown its usefulness in real-life consumer videos.

In this paper, we show that the detailed analysis of a home video database under the Bayesian perspective can be further employed both for determination of better feature spaces, and for a thorough performance evaluation of video structuring algorithms. In the first place, we analyze the issue of visual similarity of home video segments as a two-class problem. In other words, how similar are video shots that belong to the same (or to a different) event? Using multiple shot features and similarity measures, visual features are selected in order to minimize the empirical probability of segment misclassification. In the second place, we analyze the temporal structure of home video clusters, and select temporal features that reflect the patterns existing in both shots and clusters. To the best of our knowledge, there are no reported studies of feature selection for home video structuring as we have defined it. In the third place, we perform a detailed evaluation of the performance of our methodology on a ten-hour video database, with respect to cluster detection, individual shot-cluster labeling, and the effect of prior selection.

The paper is organized as follows. Section 2 describes the video structuring algorithm. Section 3 describes the procedures for feature extraction and selection. Section 4 presents the evaluation of our methodology. Section 5 draws conclusions.

2 Our approach

Hierarchical agglomerative clustering [3] has been previously used in video analysis [13]. In view of the difficulty of defining a generic generative model for intra-cluster features in home videos, we proposed to view hierarchical clustering as a *sequential binary classifier*, which at each step selects a pair of video segments s_i and s_j. and decides whether they should be merged according to Bayesian decision theory [3]. Let $\mathcal{E}$ a binary r.v. that indicates whether any pair of segments correspond to the same cluster. The Maximum a Posteriori (MAP) criterion establishes that given a realization x_{ij} of X (representing features extracted from s_i and s_j), and some knowledge about the world $\mathcal{I}$, the class $\mathcal{E}$ that must be selected is

$$\mathcal{E}^* = \arg\max_{\mathcal{E}} \Pr(\mathcal{E}|x, \mathcal{I}). \tag{1}$$

Applying Bayes' rule, the MAP criterion is expressed by

$$L = \frac{p(x|\mathcal{E}=1,\mathcal{I})\Pr(\mathcal{E}=1|\mathcal{I})}{p(x|\mathcal{E}=0,\mathcal{I})\Pr(\mathcal{E}=0|\mathcal{I})} \underset{H_0}{\overset{H_1}{\gtrless}} 1, \tag{2}$$

where $p(x|\mathcal{E},\mathcal{I})$ are the class-conditional pdfs of the observed features, $\Pr(\mathcal{E}|\mathcal{I})$ is the prior of $\mathcal{E}$, L denotes the posterior odds, H_1 denotes the hypothesis that the pair of segments belong to the same cluster and therefore should be merged, and H_0 denotes the opposite. The prior allows for the introduction of knowledge about the characteristics of home video. After performing shot boundary detection, our method starts by treating each video shot as a cluster, successively evaluates the pair of clusters that correspond to the largest L, merges only when $L \geq 1$, and continues until H_1 in Eq. 2 is not longer valid for any pair of clusters. The algorithm does not require any ad-hoc parameter determination, and generalizes previous time-constrained clustering algorithms [13].

In [4], the likelihoods are represented by Gaussian mixture models (GMMs) of global segment visual similarity, temporal adjacency and duration. However, the selection of better features would improve performance. A study of the cluster structure of the database, and of the discriminative power of features and similarity measures is presented in the next section.

3 Feature Extraction and Selection

Our data set consists of 30 MPEG-1 video clips, collected from eleven subjects, and about ten hours long (801 shots). Each sequence has a duration of around 20 minutes, and depicts typical indoor and outdoor scenarios. A third-party ground-truth at the shot and cluster levels was manually generated.

3.1 Extraction of Visual Features

Home video shots usually contain more than one appearance, due to the typical hand-held camera motion. In [4], we represented a shot by its mean color histogram. In this paper, we have adopted an approach that (1) detects *subshots* inside each shot, which approximately correspond to an individual scene appearance, and (2) extracts features from a set of *random* frames in each subshot. This generates a three-level hierarchy, as shown in Fig. 1, where a shot s_i consists of K subshots s_{ik}, $s_i = \{s_{ik}\}$, and each subshot is represented by M random frames s_{ikm}, $s_{ik} = \{s_{ikm}\}$. Additionally, we selected *joint histograms* [9] to represent color and scene structure features of each random frame. Investigated features included color in the RGB and HSV spaces, color ratios [1], edge density and edge directions [12].

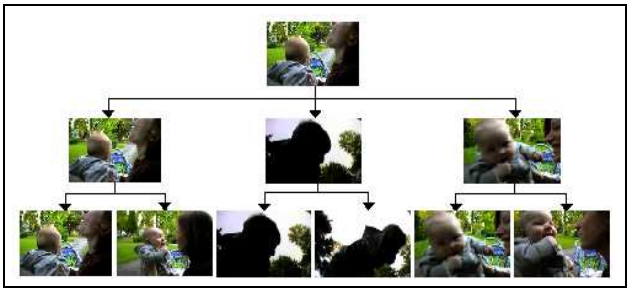

Figure 1: Intrashot hierarchy

The similarity between two shots s_i and s_j, can then be computed by $d(s_i, s_j) = \min\{d_{SS}(s_{ik}, s_{jl})\}$. In turn, the similarity between subshots s_{ik}, s_{jl}, whose random frames are represented by joint histograms h_{ikm}, h_{jln}, is defined as $d_{SS}(s_{ik}, s_{jl}) = \min\{d_\phi(h_{ikm}, h_{jln})\}$, where d_ϕ was chosen among the L_1 metric d_{L_1}, the metric based on Bhattacharyya coefficient d_B [2], and a measure based on correlation coefficient d_C.

3.2 Selection of Visual Features

We are interested in knowing how similar video shots that belong to the same (or to a different) event are. For this purpose, we estimated the distributions of intra- and inter-cluster visual similiarity, $p(d|\mathcal{E} = 1, \mathcal{I})$ and $p(d|\mathcal{E} = 0, \mathcal{I})$, for all the features and metrics just described [11]. The empirical probability of error, assuming noninformative priors, can be computed by

$$\Pr(e|\mathcal{I}) = \frac{1}{2}(\Pr(e|\mathcal{E}=0,\mathcal{I}) + \Pr(e|\mathcal{E}=1,\mathcal{I})), \tag{3}$$

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. Table 1 and Fig. 2(a-b) summarize the results.

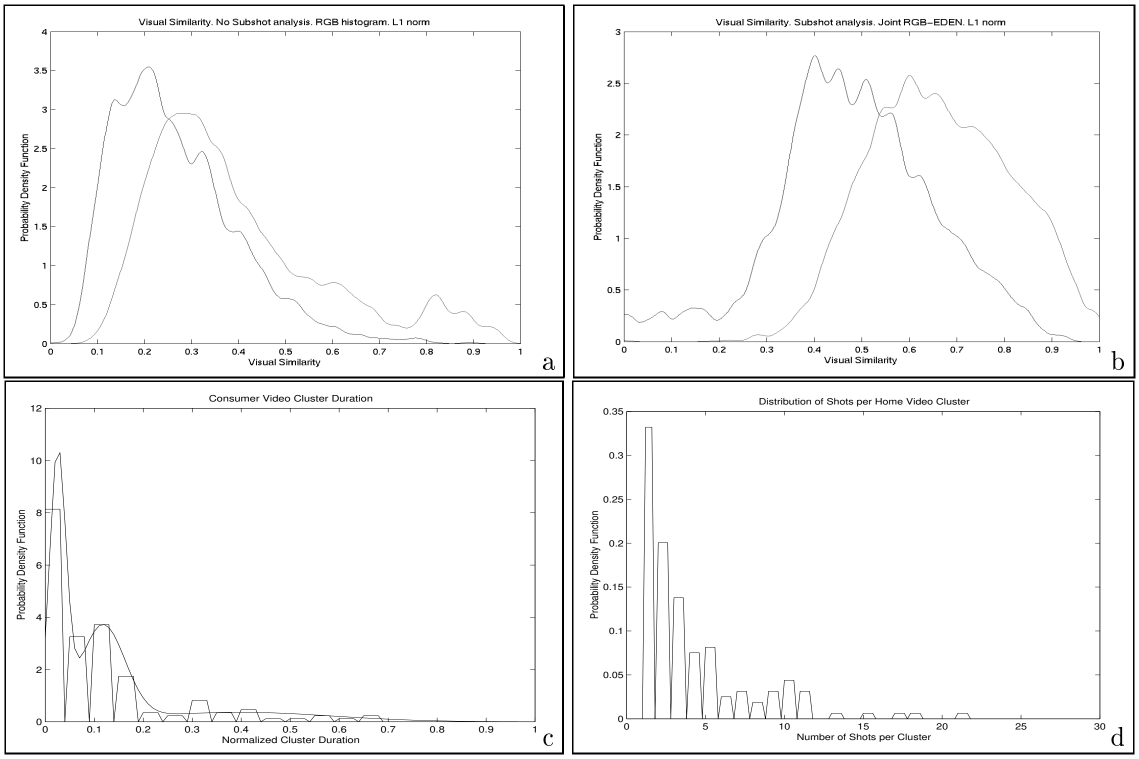

Figure 2: (a-b) Cluster visual similarity. (a) mean RGB histogram, (b) subshot analysis, RGB-edge density. Intra- and inter-cluster pdfs are denoted by continuous and dotted lines. (c-d) Temporal features. (d) Empirical pdf of normalized cluster duration, and fitted GMM (maximum duration: 1217 s). (d) Empirical pdf of number of shots per cluster. 50.3% (resp. 80.4%) of the clusters are composed of two (resp. six) or less shots.

The advantage of using subshot detection and random frames (SS+RF) as opposed to averaged shot information is shown in Fig. 2, as the former has increased the separability between the two classes. Table 1 shows the empirical probability of error computed for RGB histograms with and without subshot detection, and for 4-D 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 HSV color models. RGB-EDEN produced slightly better results than other 4-D histograms. Additionally, d_{L_1} and d_B produced better results than the correlation coefficient. The Bhattacharyya coefficient can be interpreted as the cosine of the angle between two component-wise square-rooted pdfs [2], so d_{L_1} and d_B can be thought of as representating magnitude and angle, and constitute the visual features in the clustering algorithm.

3.3 Selection of Temporal Features

Two features are typical in home videos: (1) people can focus their attention on what they record for a *limited amount of time* [6], and (2) there exists *continuity* when recording portions of *the same event.*

Indeed, the analysis of our database confirmed that (1) shot duration, cluster duration, and number of shots per cluster all present definite patterns, and (2) clusters are *localized* in time, so strong temporal adjacency can be exploited (pdfs not shown due to space reasons). Fig. 2(c-d) illustrates the patterns of temporal features. The accumulated length of two individual segments is an indication about their belonging to the same cluster (segments of increasing length become less likely to belong to the same video cluster), and is defined as

$$\Delta_{ij} = min\{|e_j - b_i|, |e_i - b_j|\} \quad (4)$$

where b_i and e_i denote the first and last frame number of s_i. A feature vector is then defined by $X = (d_{L1}, d_B, \Delta)$.

$JointHist$	Type	Pr(e)
RGB	Shots Only	0.364
RGB	SS + RF	0.319
RGB-YR	SS + RF	0.295
RGB-EDIR	SS + RF	0.286
HSV-EDIR	SS + RF	0.284
RGB-EDEN	SS + RF	0.280

Table 1: Feature Selection. L_1 metric

4 Performance Evaluation

The criteria for evaluation are $\mathcal{C}_1$: determination of the number of clusters, and $\mathcal{C}_2$: determination of the cluster label for each shot, compared to the ground-truth [3]. Although many algorithms for video clustering (home video or otherwise) have been proposed [13], [10], [7], their performance using the two criteria is unknown in several cases. We are not aware of any comparative study of video clustering techniques.

Results were generated with the leave-one-out method: one sequence was held for evaluation while all the remaining were included in the training set for density estimation [3]. Given NC, the number of clusters in the ground-truth (either in an individual sequence or in the whole database), $\mathcal{C}_1$ is evaluated by defining three variables: *Detected Clusters* (DC), *False Positives* (FP), and *False Negatives* (FN). To evaluate $\mathcal{C}_2$, *Shots In Error* (SIE) denotes 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 analyze the performance measures as probabilities (denoted in the following by lowercase symbols) using two typical estimates: the *macro-average*, which is the sample mean computed over the whole database, and the *micro-average*, in which the measure 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 its number of shots or clusters.

Table 2 evaluates the capability of our methodology to detect clusters. This is a hard problem, due to the large variability in the number of clusters in home video. Macro-averages are over-optimistic estimates as false positives in some sequences compensate for false negatives in others. In contrast, micro-averages constitute reliable measurements. The estimated value for dc was 0.75 (the ground-truth would produce a value of one). Furthermore, our method has a tendency to oversegmentation (compare fp and fn). Similar trends have been reported for other types of video [13], [10]. A detailed analysis indicated that many false negatives consist of only one or two shots. As a baseline, we show the poor result that is obtained with an algorithm that randomly estimates the number of clusters for

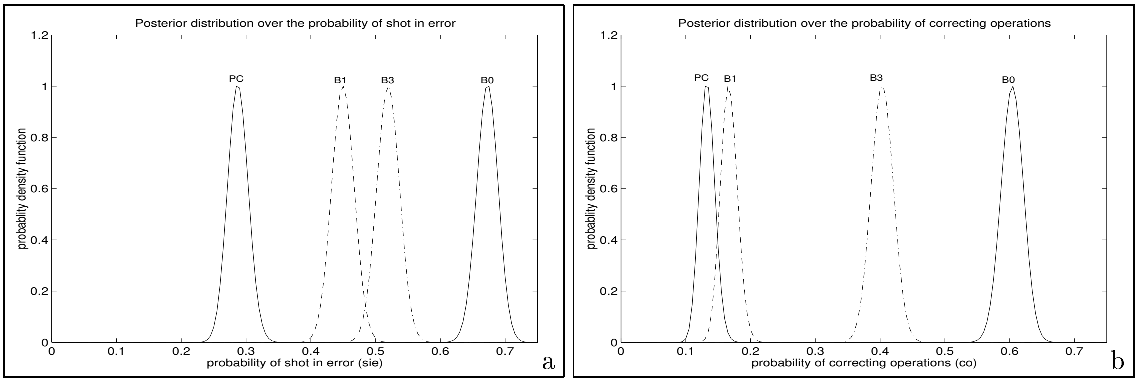

Figure 3: (a) Posteriors of the probability of shot in error for different structuring algorithms. PC denotes our approach. (c) Posteriors of the probability of correcting operations.

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.

Method	dc_m	fp_m	fn_m
Random Clustering	0.470	0.514	0.015
Probabilistic Clustering	0.750	0.171	0.079

Table 2: Cluster Detection Performance

Table 3 describes the performance in terms of shot-cluster assignment, for which both macro- and micro-averages are useful measurements (the ground-truth generates a zero value for all cases). Differences between them indicate variation of individual performance. We selected a number of baseline methods for comparison, assuming the *correct* number of clusters for each sequence: B_1, which assigns a uniform number of shots per cluster; (ii) B_2, a version of K-means, in which the centroids were initialized with randomly selected shots from each sequence; and (iii) B_3, K-means, with uniform initialization. We also present results for B_0 (random clustering). Our methodology outperformed all of the baseline methods. Using macro-averages (resp. micro-averages) as measurement, our methodology assigned 71.1 (resp. 71.4)% of the shots to the correct cluster. Interestingly, uniform shot-assigment performed better than K-means. 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 to correct the cluster assignments in a 20-minute video.

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 3: Shot Assignment Performance

Using the Bayesian approach [3], suppose we observe n shots in error out of N. The likelihood function $\Pr(n|sie)$ is a binomial distribution. Assuming a uniform prior, the expression for the posterior is $p(sie|n) \propto sie^n(1-sie)^{N-n}$. Fig. 3(a) compares the posteriors over the probability of shot in error, estimated for the different clustering methods. Fig. 3(b) presents the corresponding analysis for the posterior of the probability of correcting operations $p(co|n)$.

The effect of the prior distribution is shown in Table 4. A uniform prior does not make use 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, but excessive merging affected performance. On the other hand, the ML-estimated prior for our database is $\Pr(\mathcal{E}|\mathcal{I}) = \{0.87, 0.13\}$.

Method	dc_m	fp_m	fn_m	sie_m	sie_M
Uniform prior	0.573	0.000	0.427	0.309	0.393
Empirical prior	0.750	0.171	0.079	0.286	0.289

Table 4: Effect of Prior Probability

An example of the generated video structure in a tree fashion is shown in Fig.4. Each shot is displayed as a column of random frames. Qualitatively, our methodology provides quite good results. Fig. 5 shows typical merging errors. As a general trend, outdoor clusters are harder to segment. There are three reasons for erroneous merging: (1) high visual similarity between semantically disjoint but temporally adjacent video clusters, (2) shots of very short duration, and (3) clusters of very short duration. The reasons for oversegmentation of clusters are (1) high intra-cluster visual variability, and (2) unusually long clusters.

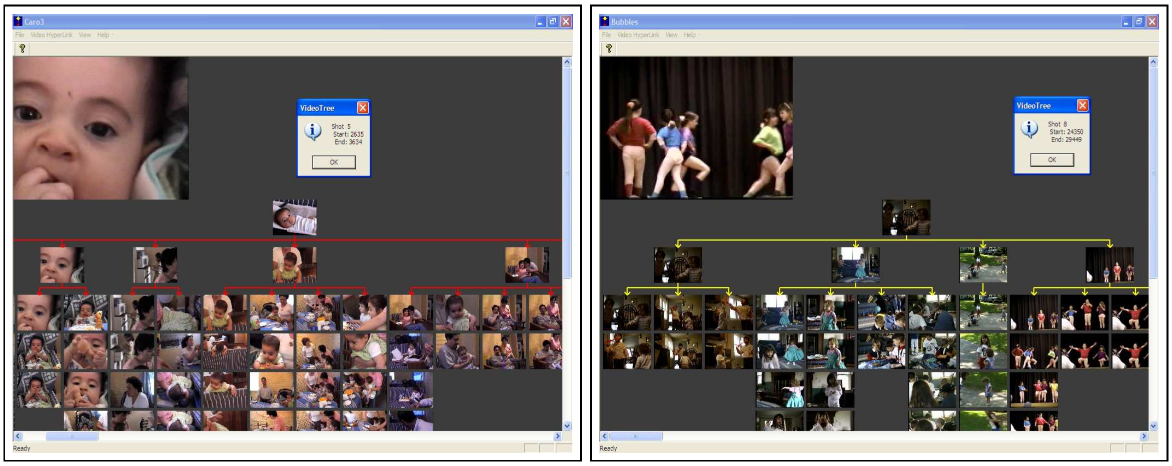

Figure 4: Examples of video structuring on two *Family* video sequences (detail). The root node corresponds to the video sequence, the middle nodes to the clusters, and the leaves to random frames from shots.

Figure 5: (a-b), (c-d) Frames extracted from pairs of video shots that were erroneously merged by our methodology.

5 Concluding Remarks

A detailed analysis of the visual and temporal structure of a relatively large home video database offered a number of clues for probabilistic video structuring. The obtained results are encouraging, but also illustrate the complexity of the problem at hand. In particular, to quantify judgement differences between people, due to the uncertainty of the contents, an alternative for performance evaluation could consist in the definition of a pdf of human judgment in a Bayesian context, and its use to quantify automatic algorithms. This approach is under evaluation.

Acknowledgements. Several of the analyzed video sequences belong to the Eastman Kodak 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] D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift," in *Proc. IEEE CVPR.*, Hilton Head Island, S.C., June 2000.

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

[4] D. Gatica-Perez, M.-T. Sun, and A. Loui, "Consumer Video Structuring by Probabilistic Merging of Video Segments," in *Proc. IEEE ICME*, Tokyo, Aug. 2001.

[5] G. Iyengar, and A. Lippman, "Content-based browsing and edition of unstructured video," in *Proc. IEEE ICME*, New York City, Aug. 2000.

[6] J.R. Kender and Yeo, B.L., "On the Structure and Analysis of Home Videos," in *Proc. ACCV*, Taipei, Jan. 2000.

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

[8] 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.

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

[10] 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.

[11] D.W. Scott, *Multivariate Density Estimation*, Wiley, 1992.

[12] 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.

[13] 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.