A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval

Milind Ramesh Naphade and Thomas S. Huang, *Fellow, IEEE*

***Abstract*—Semantic filtering and retrieval of multimedia content is crucial for efficient use of the multimedia data repositories. Video query by semantic keywords is one of the most difficult problems in multimedia data retrieval. The difficulty lies in the mapping between low-level video representation and high-level semantics. We therefore formulate the multimedia content access problem as a multimedia pattern recognition problem. We propose a probabilistic framework for semantic video indexing, which can support filtering and retrieval and facilitate efficient content-based access. To map low-level features to high-level semantics we propose probabilistic multimedia objects (*multijects*). Examples of multijects in movies include *explosion, mountain, beach, outdoor, music* etc. Semantic concepts in videos interact and to model this interaction explicitly, we propose a network of multijects (*multinet*). Using probabilistic models for six site multijects, *rocks, sky, snow, water-body, forestry/greenery* and *outdoor* and using a Bayesian belief network as the multinet we demonstrate the application of this framework to semantic indexing. We demonstrate how detection performance can be significantly improved using the multinet to take interconceptual relationships into account. We also show how the multinet can fuse heterogeneous features to support detection based on inference and reasoning.**

***Index Terms*—Bayesian belief networks, hidden Markov models, likelihood ratio test, multimedia understanding, probabilistic graphical networks, ROC curves, semantic video indexing, query by example, query by keywords, semantic video indexing.**

I. Introduction

GENERATION of digital multimedia content has increased tremendously in recent years. Rapid advances in the technology for media capture, storage and transmission and increasingly affordable prices of these devices has contributed to an amazing growth in the amount of multimedia, that is available on the internet. Whether it is sharing of picture albums and home videos, advertising of movies through interactive preview clips, live broadcasts of various shows or multimedia reports of news as it happens, multimedia information has found in the internet, a medium to reach us. Add to that, innovations in hand-held and portable computing devices and wired and wireless communication technology (pocket PCs, organizers, cell phones) on one end and broadband internet devices on the other and we have a huge supply and dissemination of unclassified multimedia information flooding us.

As content generation and dissemination grows explosively, the need for tools to filter, classify, search and retrieve this content efficiently becomes even more acute. Tools for information retrieval in text databases cannot be extended to this problem. Lack of tools for efficient access and data-mining threatens to render most of this data useless. Apart from a few exceptions [1] most state of the art video retrieval systems neglect the audio component and support the paradigm of visual query by example using similarity in low-level media features. Examples include [2]–[6] etc. In this paradigm, the query must be phrased in terms of a video clip or at least a few key-frames extracted from the query clip. The retrieval is based on a matching algorithm, which ranks the target clips according to a heuristic measure of similarity between the query and the target. This approach is suited for browsing and low-level search, but has limitations. It is unrealistic to expect that the user has access to one or more representative clips. Also, high-level similarity may not correspond to low-level feature based similarity if there is no attempt to understand the semantics of the query.

To address the aforementioned problems, we need a semantic indexing, filtering and retrieval scheme. For a system to fetch clips of an *explosion on a beach*, the system must understand how the concepts *explosion* and *beach* are represented. This is a very difficult problem. The difficulty lies in the gap, that exists between low-level media features and high-level semantics. In this paper we present a novel probabilistic framework to bridge this gap to some extent. We view the problem of semantic video indexing as a multimedia understanding problem. We apply advanced statistical pattern recognition and learning techniques to develop generic models representing semantic concepts.

Semantic concepts do not occur in isolation. There is always a context to the co-occurrence of semantic concepts in a video scene. We believe that it is beneficial to model this context. We use a probabilistic graphical network to model this context and demonstrate how this leads to a significant improvement in detection performance. We also show how the context can be used to infer about some concepts based on their relation with other detected concepts. We develop models for the following semantic concepts: *sky, snow, rocky-terrain, water-body* and *forestry/greenery*, and *outdoor*. Using these concepts in our experiments, we demonstrate how filtering and key-word based retrieval can be performed on multimedia databases.

The paper is organized as follows. We review existing techniques in content-based video retrieval in Section II. We present the probabilistic framework in Section III. Preprocessing, feature extraction and representation are discussed in Section IV. We discuss the actual process of developing models for semantic concepts in Section V. The probabilistic graphical network used

Manuscript received May 31, 2000; revised December 1, 2000. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jenq-Neng Hwang.

The authors are with the Department of Electrical and Computer Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (e-mail: milind@ifp.uiuc.edu; huang@ifp.uiuc.edu).

Publisher Item Identifier S 1520-9210(01)01865-X.

to model the context is described in Section VI. We show how the use of the multinet enhances detection performance in Section VII. Fusion of soft-decisions from heterogeneous classifiers using the multinet is shown in Section VIII. We then discuss how the probabilistic models can be used for filtering and retrieval in Section IX. Finally, directions for future research and conclusions are presented in Section X.

II. Review

The first step in any video data management system is the segmentation of the video track into smaller units. Segmentation of video into shots is well understood. Processing for segmentation can be done in compressed [7]–[10] as well as uncompressed domain [11]. Shots can be grouped based on continuity, temporal proximity and similarity to form scenes. Also, keyframes can be extracted from shots to help browsing. A typical structure imposed on the videos for efficient browsing is shown in Fig. 1.

In addition to efficient browsing, video data management also demands tools for efficient access based on some paradigm of query. The most supported paradigm of query is the paradigm of query by example (QBE). Systems that support this paradigm include [2], [4], [3], [6], [5]. This is just an extension of the technology for querying image databases, using one or more set of images or sketches. In its most powerful form, the extension to video includes use of an object-based model [2], which permits a sketch of objects of specific color, size, location, and motion trajectory. In simpler forms it supports queries in terms of video clips, which are matched with clips in the database [6]. An inherent assumption is that the user knows (perfectly or reasonably) the size, shape and/or motion trajectory of the object, being searched for in the video [2] or has sufficient number of representative clips [6], or keyframes [4] to pose the query. All such systems use low-level features based on color, texture, motion and shape to perform matching of video clips in one way or another and then rank the retrieved clips in terms of some similarity measure.

While there have been tremendous advances in speech recognition and speech processing related technologies, there has been little progress in terms of management of nonspeech audio data. Most of the indexing and retrieval in audio assumes human speech (sitcoms, radio interviews, news) with relatively noise-free environment and works on a vocabulary of words. Recent examples include [12]. Recently, there have been attempts to segment the sound-track in motion pictures [13], [14] and television comedies [15].

Query using keywords, which represent semantic concepts has motivated recent research in semantic video indexing [16]–[18]. Recent attempts to introduce semantics in the structuring of videos includes [14], [19], [20]. We [16] present novel ideas in semantic video indexing by learning probabilistic multimedia representations of semantic concepts including semantic events like *explosion* and sites like *waterfall*. Chang *et al.* [17] introduce the notion of semantic visual templates. Wolf *et al.* use hidden Markov models to parse video [19]. Ferman *et al.* [20] attempt to model semantic structures like *dialogues* in video. It is increasingly obvious, that efficient filtering and retrieval cannot be facilitated, unless the semantics of multimedia content is addressed.

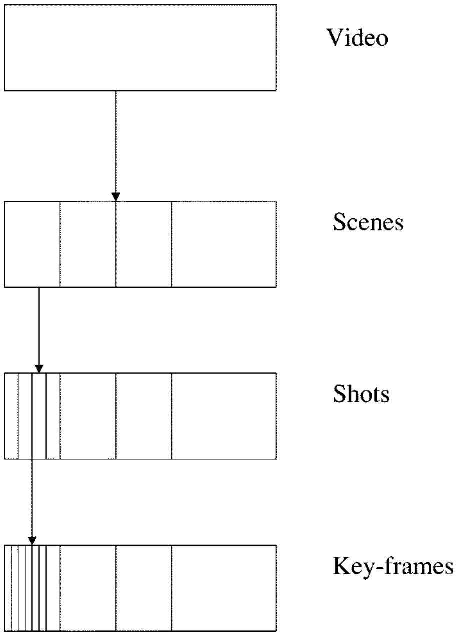

Fig. 1. Common structure imposed on a video. The hierarchy of scenes, shots and key-frames is useful in efficient browsing.

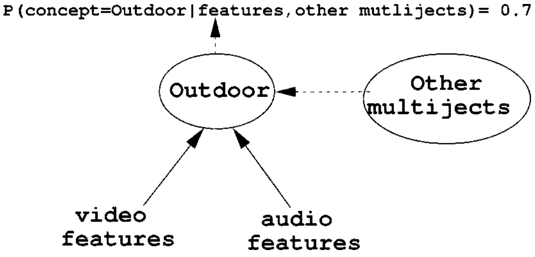

Fig. 2. Multiject for the semantic concept *outdoor*. The media support for the label *outdoor* is in the form of audio-visual features. In addition to this there may be support from other multijects representing semantic concepts like *sky*.

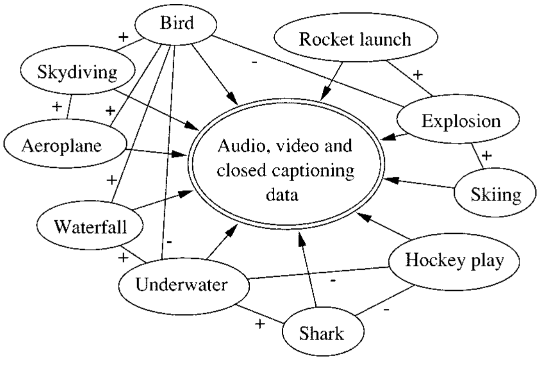

Fig. 3. Conceptual figure of a multinet. The multinet captures the interaction between various semantic concepts. The edges between multijects denote interaction and the signs on the edges denote the nature of interaction.

III. Framework

Users of multimedia (audio-visual) databases are interested in finding video clips using queries, which represent high-level concepts. While such semantic queries are very difficult to support exhaustively, they might be supported partially, if models representing semantic concepts are available. User queries might involve *sky, car, mountain, sunset, beach, explosion*, etc. Detection of some of these concepts may be possible, while some others may be difficult to model. To support such queries, we define a *multiject*. A multiject [16] is a probabilistic multimedia object, which has a semantic label and which summarizes a time sequence of features extracted from multiple media. *Multijects* can belong to any of the three categories: objects (*car, man, helicopter*), sites (*outdoor, beach*), or events (*explosion, man-walking, ball-game*). The features themselves may be low-level features, intermediate level visual templates or specialized concept detectors like face-detectors or multijects representing other semantic concepts. Fig. 2 shows an example.

Semantic concepts are related to each other. One of the main contributions of this paper is a probabilistic graphical framework to model this interaction or context. It is intuitively obvious that detection of certain multijects boosts the chances of detecting certain other multijects. Similarly, some multijects are less likely to occur in the presence of others. For example, the detection of *sky* and *water* boosts the chances of *beach*, and reduces the chances of detecting *Indoor*. An important observation from this interaction is that it might be possible to infer some concepts (whose detection may be difficult) based on their interaction with other concepts (which are relatively easier to detect). For example, it may be possible to detect human speech in the audio stream and detect a human face in the video stream and infer the concept *human talking*. To integrate all multijects and model their interaction explicitly we therefore propose a network of multijects, which we call a *multinet*. A conceptual figure of a multinet is shown in Fig. 3 with the positive signs in the figure indicating a positive interaction and negative signs indicating a negative interaction.

By modeling the relationship between multijects we can

Enhance detection: The use of mutual information can enhance detection of multijects.

Support inference: Some multijects may not provide us with the required degree of invariance in feature-spaces. For the detection of such multijects, the multinet can support inference based on the interaction of these multijects with other multijects (which can be detected with greater ease). For example, we can detect the multiject *beach* based on the presence of such multijects as *water, sand, trees, boat*. Based on this detection of *beach* we can then claim that the scene is an *outdoor* scene.

Impose prior knowledge: The multinet can provide the mechanism for imposing time-varying or time-invariant prior knowledge of multiple modalities and enforce context-changes on the structure. For example, knowledge that a movie is an action movie, may be used to increase the prior probabilities of such multijects as *gunshots, explosion*.

IV. Preprocessing and Feature Extraction

Each video frame can be segmented into regions. Within each video shot, these regions evolve over time along with the audio track to convey a meaningful story-line or narrative. In terms of visual presence, some semantic concepts exist regionally (*face*). For many a concept no single region in the video frame is sufficient to convey the concept and the meaning is conveyed only by all the regions in the video frame. When this happens, we say, these concepts occur globally (e.g., *outdoor*). The multimodal support for a concept may thus imply association of one or more regions with the audio track. To build multiject models we thus need to extract features at regional and global level from the visual stream and features from the audio stream as well.

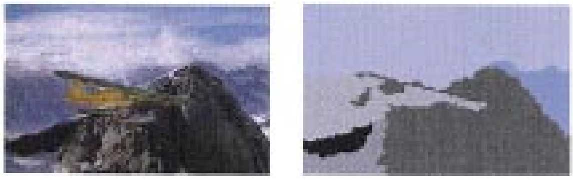

Fig. 4. Segmentation.

A. Preprocessing the Video Track

The video clips are segmented into shots using the algorithm in [11]. We then use the spatio-temporal segmentation in [2][1] applied separately to each shot to obtain regions homogeneous in color and motion. Depending on the nature of the movie and the story-line, shots may range from a few frames to a few hundred frames. For large shots, artificial cuts are introduced every 2 s. This ensures, that the spatio-temporal tracking and segmentation does not break down due to considerable appearance and disappearance of regions. The segmentation and tracking algorithm uses color, edge and motion to perform segmentation and computes the optical flow for motion estimation. Segmented regions are then merged using morphological operations and based on coherent motion and weak edges. Fig. 4 shows a video frame and its segmented version with six dominant segments.

The segmentation algorithm is tuned to obtain large blobs. The idea is to prohibit a single concept to be broken down into multiple regions as far as possible. The imperfections thus mostly result in multiple concepts existing in a single region. In Section V-C we discuss how the system is made tolerant to such imperfections to a large extent.

The segments thus obtained are labeled manually to create the ground truth. Since they are tracked within each shot using optical flow, the labels can be propagated to instances of the segments in all the frames within the shot.

Each region is then processed to extract a set of features characterizing the visual properties including the color, texture, motion and structure of the region. We extract the following set of features.[2]

[1] The authors thank D. Zhong and Dr. S. F. Chang for the spatio-temporal segmentation and tracking algorithm.

[2] Our aim is to work with a set of reasonable features. There is no claim to the optimality of this set of features. Better features will obviously lead to better performance. Also dimensionality reduction is possible and even desirable when the number of training samples is small but we will not focus on those issues.

Fig. 5. Collection of shots from some of the movies in the database.

TABLE I
MAXIMUM LIKELIHOOD BINARY CLASSIFICATION PERFORMANCE OVER SEGMENTED REGIONS FOR *SITE* MULTIJECTS USING GAUSSIAN MIXTURE CLASS CONDITIONAL DENSITY FUNCTIONS FOR THE TRUE AND NULL HYPOTHESES FOR EACH MULTIJECT

multiject	Detection Accuracy	False Alarm
rocks	77%	24.1%
sky	81.8%	11.9%
snow	81.5%	12.9%
water	79.4%	15.6%
forest	85.1%	14.9%
Overall	80.96%	15.88%

Color: A normalized, linearized[3] three-channel HSV histogram is used, with 12 bins each, for hue (H), saturation (S) and intensity (V). The invariance to size, shape, intra-frame motion and their relative insensitivity to noise makes color histograms the most popular features for color content description.

Texture: Texture is a spatial property. A two dimensional dependence matrix, which captures the spatial dependence of gray-level values contributing to the perception of texture is called a gray-level co-occurrence matrix (GLCM). A GLCM is a statistical measure extensively used in texture analysis. In general for pairs of intensity values $(i,\, j)$, we denote

$$p(i,\, j,\, d,\, \theta) = \frac{P(i,\, j,\, d,\, \theta)}{N(d, \theta)} \tag{1}$$

where $P(\bullet)$ is the GLCM for the displacement vector d and orientation θ and $N(\bullet)$ is the normalizing factor to make the left hand side of (1) a probability distribution. In our work, we compute GLCMs of the V channel using 32 gray-levels and at four orientations: corresponding to θ values of $0°$, $45°$, $90°$ and $135°$ degrees respectively. For all four GLCMs we consider pixels, which are at a distance of one unit from the current pixel respectively ($d = 1$). For each of the four matrices (corresponding to a fixed d and a θ), six statistical features of the GLCMs are computed. The features are Contrast, Energy, Entropy, Homogeneity, Correlation, and Inverse Difference Moment [21].

Structure: To capture the structure within each region, a Sobel operator with a 3×3 window is applied to each region and the edge map is obtained. Using this edge map an 18-bin histogram of edge directions is obtained as in [22]. The edge

[3] A linearized histogram of multiple channels is obtained by concatenating the histogram of each channel. This avoids dealing with multi-dimensional histograms.

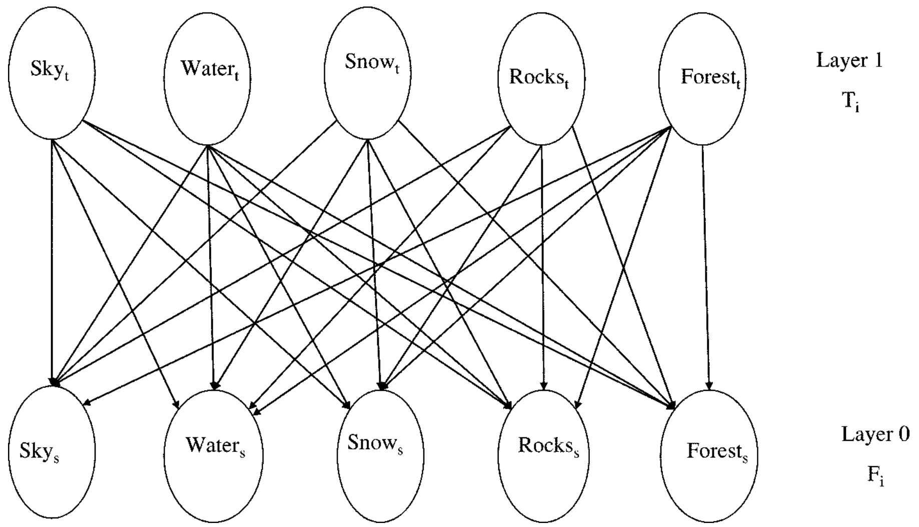

Fig. 6. Two-layered Bayesian multinet. All nodes in the network denote binary random variables. Nodes in Layer 1 are the parent nodes representing the true distributions. Nodes in Layer 0 are the random variables of Equation (6) representing the frame-level multiject-based semantic features. The soft decisions for nodes at Layer 0 are obtained using Equation (6). The soft decisions for nodes in Layer 1 are obtained through inference using the parameterized multinet and by introducing evidence at the nodes in Layer 0.

direction histogram is supposed to be a robust representation of structure [23].

Motion: The interframe affine motion parameters for each region tracked by the spatio-temporal segmentation algorithm are used as motion features.

Color Moments: The first order moments and the second order central moments are computed for each of the three channels H, S, and V. In all, 98 features are extracted to represent the visual properties of the region, of which 84 features (color, texture, structure and moments) are used for sites. For objects and events, all 98 features are used. A similar set of features can also be obtained at the global level without segmentation and also on difference frames obtained using successive consecutive frames [16].

B. Preprocessing the Soundtrack

The soundtrack is digitized at 44.1 kHz sampling frequency. It is then preprocessed using a 20 channel filter-bank. Cepstral transformation gives 15 mel frequency cepstral coefficients (MFCC), 15 delta, and two energy coefficients. A window width of 25 ms and overlap of 10 ms is used. This gives a 32-coefficients feature vector per window. The segmentation of the audio-track and the detection of audio-multijects is integrated. Since we use motion picture soundtracks, we cannot restrict the type of possible audio signal sources to speech only. The most common sources are human speech, music, animal sounds (*horse, bear*), mechanical sounds (*automobile engine, helicopter*) and natural sounds (*wind, water, thunder, waves*). The problem with the motion picture soundtrack is that usually, no single source contributes individually at any time. Usually, there are multiple sound sources present in the track. This complicates the analysis of the audio track. We attempt to develop models, which account for the co-occurrence of multiple sound sources and use this to perform integrated segmentation and multiject-detection of the sound track [13]. Results about audio-multijects like *music, human-speech* etc. are not presented here and can be found in [13].

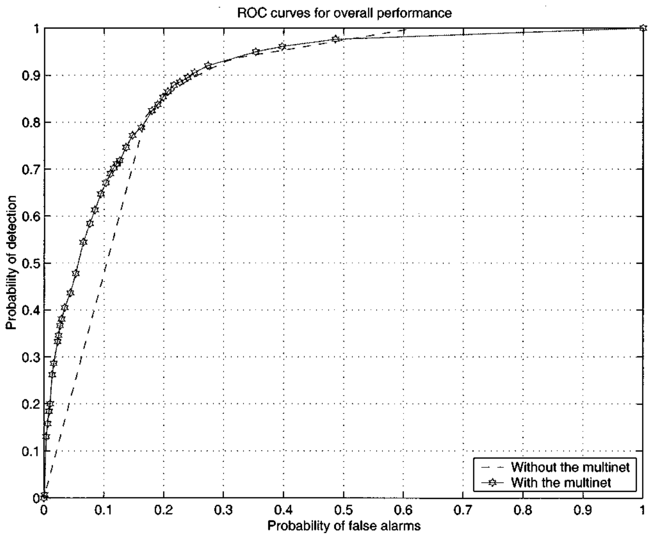

Fig. 7. ROC curves for overall performance using a Bayesian multinet with soft child nodes and binary parent nodes. The *OR* operator is used for frame-level fusion. The multiject-based ROC curve corresponds to classification by a likelihood ratio test using the soft decisions of (6). The multinet-based ROC curve corresponds to the likelihood ratio test using soft decisions obtained at the nodes in Layer 1 of Fig. 6. Clearly, the multinet enhances detection performance.

V. Estimating Multiject Models

An identical approach is used to model concepts in video and/or audio.

A. Multijects Based on Video

Let $\vec{X}_j$ be the feature vector for region j. We define two hypotheses H_0 and H_1. Under each hypothesis, we assume that the feature vector is drawn from a distinct probability distribution as defined in (2):

$$\begin{aligned} H_0&: \vec{X}_j \sim P_0\left(\vec{X}_j\right) \\ H_1&: \vec{X}_j \sim P_1\left(\vec{X}_j\right). \end{aligned} \qquad (2)$$

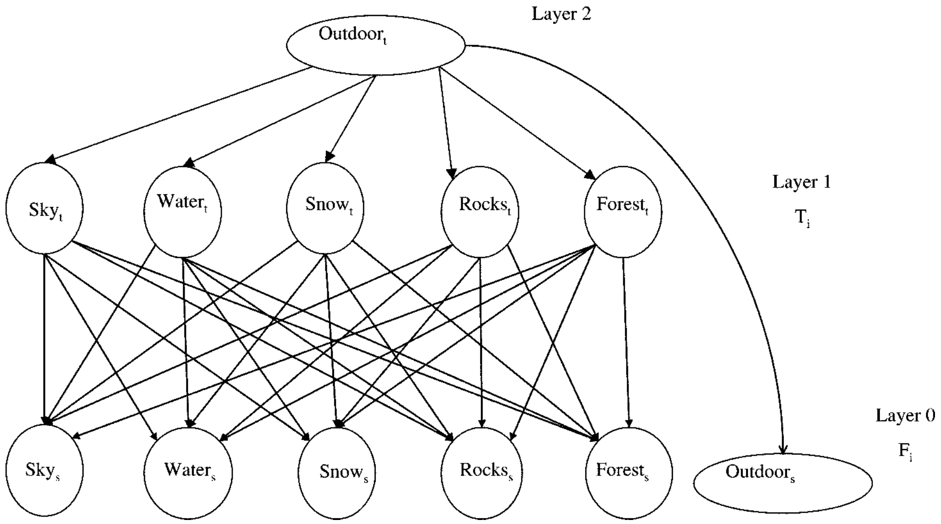

Fig. 8. Modification to the Bayesian multinet in Fig. 6. The node $outdoor_s$ represents the global multiject existing at the frame level. The node $outdoor_t$ in Layer 2 uses the soft decisions of the $outdoor_s$ node in addition to the soft decisions of the five nodes in Layer 1 to make a soft decision about the presence of the *outdoor* multiject.

$P_0(\vec{X}_j)$ and $P_1(\vec{X}_j)$ denote the class conditional probability density functions conditioned on the null hypothesis (concept absent) and the true hypothesis (concept present) respectively. In case of sites the class conditional density functions of the feature vector under the true and null hypotheses are modeled as mixture of diagonal Gaussian components (GMM). The temporal flow is not taken into consideration. In case of objects and events we use hidden Markov models (HMM) with continuous Gaussian mixture observation densities in each state for modeling the time series of the feature vectors of all the frames within a shot under the null and true hypotheses. The EM algorithm [24] is used in both cases to estimate the parameters of the density models, i.e., the mean vectors, covariance matrices, mixing proportions (GMM and HMMS), and state transition matrices (HMM).

Here we present results of regional site multijects *sky, water, forest, rocks*, and *snow*. These multijects are based only on visual features and are used in experiments reported in the remainder of this paper.

B. Experimental Setup

We have digitized movies of different genres including action, adventure, romance, and drama to create a database of a few hours of video. Data from eight movies has been used for the experiments. Fig. 5 shows a collection of shots from some of the movies in the database and should convince the reader of the tremendous variability in the data and representative nature of the database. The MPEG streams of data are decompressed to perform shot-boundary detection, spatio-temporal video-region segmentation and tracking and subsequent feature extraction. For all the experiments reported in this paper, segments from over 1800 frames are used for training and segments from another 9400 frames are used for testing. These images are obtained by downsampling the videos temporally, in order to avoid redundant images in the training set. Each image is of the size 176×112 pixels. The number of Gaussian components in the mixtures for the null and true hypotheses can be chosen optimally based on a tradeoff between performance and number of parameters used for describing the distributions. In our experiment we heuristically use five components for the distribution under the true hypothesis and ten components for that under the null hypothesis. It can be argued that the feature distribution under the null hypothesis should ideally be uniform over the feature space. This would be true if we had infinite training data. However due to finite training data, the distribution under the null hypothesis may have multiple modes each due to a different class of negative examples. We may therefore need more modes to cover the feature space in this case than to cover the space of examples for the true hypothesis. Our heuristic choice of mixing components is based on this belief. We model the five site multijects: *rocks* representing rocky terrain, *sky* representing the sky, *snow* representing snow-covered ground, *water* representing water-bodies like lakes, rivers, oceans etc., and *forest* representing vegetation and greenery.

The detection performance of each of the five *site* multijects over the test-set based on maximum-likelihood binary classification using the GMMs for the two hypotheses is given in Table I.

C. Integrating Regional Likelihoods to Obtain Frame Level Multiject Features

A multinet models the interaction between multijects at the frame-level. To obtain frame-level features we need to fuse the region-level features. The strategy for fusing region-level multijects to obtain frame-level semantic features must take imperfections in segmentation into account. As described earlier the segmentation and tracking algorithm favors large regions with greater possibility of multiple concepts in a single region than a single concept scattered across multiple regions. We therefore check each region for each concept independently. By doing this we avoid a loss of information that could have occurred if we used classes, which were mutually exclusive and chose one class (concept) for each region. Also by fusing the information from all the dominant regions in the frame, we reduce the probability

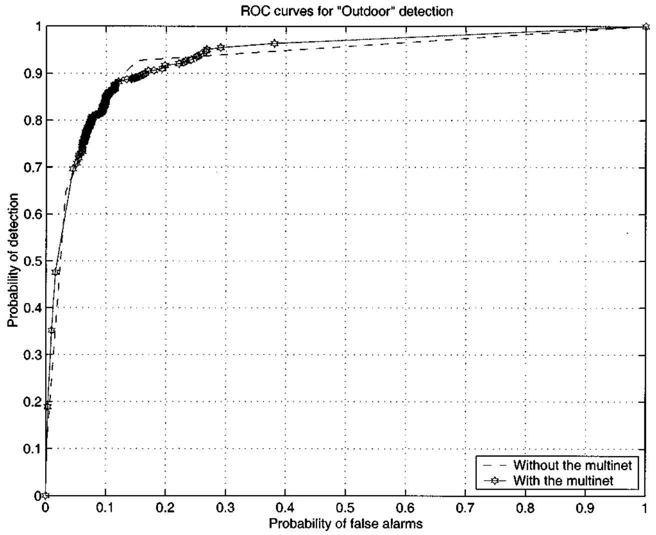

Fig. 9. ROC curves for the Bayesian multinet of Fig. 8. The curve in blue corresponds to the detection using only the Gaussian mixture models for the outdoor multiject built using global media features only [(13)]. The curve in red represents the detection of the *outdoor* multiject, which uses the soft decision of the global-features based classifier as well as the soft decisions of the five other site multijects. [(14)].

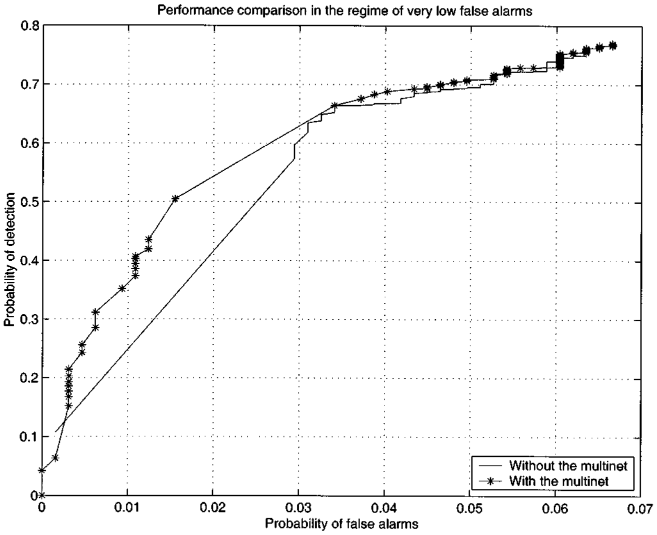

Fig. 10. Magnified portion of the ROC curves of Fig. 9 for the range $0 \leq P_f \leq 0.1$.

of misdetection at frame level compared to that at the region level. For the binary classification of each concept in each region we define binary random variables R_{ij} in (3)

$$R_{ij} = \begin{array}{ll} 1 & \text{if concept } i \text{ is present in region } j \\ 0 & \text{otherwise.} \end{array} \tag{3}$$

Assuming uniform priors on the presence or absence of any concept in any region and using Bayes' rule we then obtain [(4)]

$$P\left(R_{ij} = 1 \middle| \vec{X}_j\right) = \frac{P\left(\vec{X}_j \middle| R_{ij} = 1\right)}{P\left(\vec{X}_j \middle| R_{ij} = 1\right) + P\left(\vec{X}_j \middle| R_{ij} = 0\right)}. \tag{4}$$

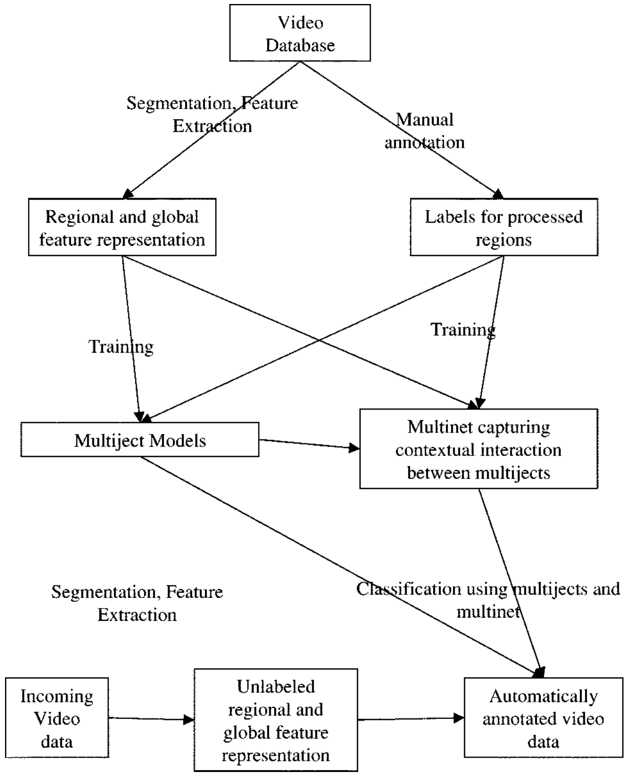
Video Database
Segmentation, Feature Extraction
Manual annotation
Regional and global feature representation
Labels for processed regions
Training
Training
Multiject Models
Multinet capturing contextual interaction between multijects
Segmentation, Feature Extraction
Classification using multijects and multinet
Incoming Video data
Unlabeled regional and global feature representation
Automatically annotated video data

Fig. 11. Block diagram of the multiject-based multimedia understanding system. Videos from the training set are spatio-temporally segmented and manually annotated. This training set is then used to estimate multiject models for various semantic concepts. It is also used to learn the relation between the concepts i.e., the context. The multiject models along with the multinet are then used for automatic annotation, filtering and retrieval of semantics.

The multijects used here are region-level detectors. For fusing regional information at frame level, we define frame-level semantic features F_i, $i \in \{1, \cdots, N\}$ (N is the number of concepts) in (5)

$$F_i = \begin{array}{ll} 1 & \text{if concept } i \text{ is present in the current frame} \\ 0 & \text{otherwise.} \end{array} \tag{5}$$

To fuse the region-level concepts we can use various functions. Let the number of regions in the frame be M. Using the compact notation $\mathcal{X} = \{\vec{X}_1, \cdots, \vec{X}_M\}$ the fusion is defined in (6)

$$\begin{aligned} P(F_i = 0|\mathcal{X}) &= \prod_{j=1}^{M} P\left(R_{ij} = 0 \middle| \vec{X}_j\right) \\ P(F_i = 1|\mathcal{X}) &= 1 - P(F_i = 0|\mathcal{X}). \end{aligned} \tag{6}$$

For multijects based on global features (representing semantic concepts at a global level) like *outdoor* the multijects exist at the frame-level and there is no need of regions-to-frame fusion.

VI. A Bayesian Multinet

To model the interaction between frame-level concepts, we propose a Bayesian belief network as the multinet. A Bayesian belief network [25]–[27] is a probabilistic graphical network, which specifies a probability distribution over a set of random variables, which are represented by the nodes of the network. Any node in a Bayesian network is independent of the rest of the nodes in the network given the values of its parents. A Bayesian net is a directed acyclic graph and the direction of the edges between the nodes can be interpreted as causality among the random variables represented by the nodes.

We now describe the Bayesian network based multinet. In Section V, each concept was detected independent of the others. The binary random variables F_i, denoting the presence/absence of multijects at frame-level have been derived from region-level features for each concept i separately. However, they represent semantic concepts in a movie. Since they convey meaning within a context, they must interact. Since the random variables F_i do not capture this interaction, we define another set of binary random variables T_i, $i \in \{1, \cdots, N\}$. For all $i \in \{1, \cdots, N\}$, T_i and F_i correspond to the same concept but T_i takes into account the dependence of concepts at frame-level which F_i ignored earlier. Fig. 6 shows the dependence between T_i and F_i for $N = 5$ in a two-layered Bayesian network. The five concepts used here are *sky, water, forest, rocks* and *snow*. In Fig. 6, parent nodes appear in Layer 1 and represent T_i the random variables denoting the true distributions of the concepts (hence we use subscript t). Child nodes appear in Layer 0 and represent the concepts F_i. The child nodes are conditionally independent given the parent nodes T_i (The subscript for child nodes s indicates isolated or stand-alone detection because this detection is done separately for each multiject).

The dependence between the various multijects at frame-level is modeled in the conditional distributions $P(F_i|T_1, \cdots, T_N)$. During the training phase[4] we first obtain the probability of each node in Layer 0 being ON and OFF using (6) [i.e., $\{P(F_i = 0|\mathcal{X}), P(F_i = 1|\mathcal{X})\}$, $i \in \{1, \cdots N\}$]. The ground truth for nodes in Layer 1 is already available. The parameters of the network are then estimated using the training set so as to maximize the likelihood of the training set given the Bayesian network. The EM algorithm is used for the likelihood maximization. Then, during the testing or *inference* phase, the soft decisions at the nodes in Layer 0 are used to feed the network and we infer the nodes in Layer 1 being ON and OFF, i.e., $P(T_i = 1|F_1, \cdots, F_N)$ and $P(T_i = 0|F_1, \cdots, F_N)$. This effectively captures the interaction between the concepts since the inference is based on the soft decisions of all the detected concepts. We will show in Section VII, that this leads to a significant improvement in detection performance.

VII. Improvement in Detection Performance

In this section, we support our claim that modeling of context can lead to a significant improvement in detection performance. To evaluate the performance of the system over the frames in the test-set, we propose to compare the detection performance using the receiver operating characteristics (ROC) curves. For a binary classification problem, an ROC curve is a parametric plot obtained by plotting the probability of detection against the probability of false alarms for different values of the parameter.

In testing the hypothesis H_0 versus the hypothesis H_1, two types of errors are possible. H_0 can be falsely rejected or H_1 can be falsely rejected. The design of a test for H_0 versus H_1 involves a trade-off, since one can always be made arbitrarily small at the expense of the other. We employ the Neyman–Pearson criterion [28] for making this tradeoff. The idea is to place a bound on the false alarm probability and then to maximize the detection rate subject to this constraint; i.e., the Neyman–Pearson design criterion is

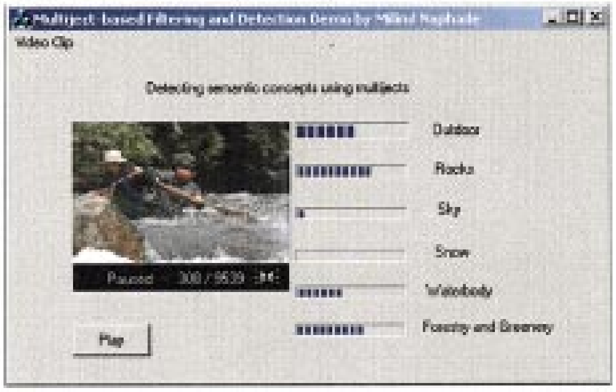
Multiject-based Filtering and Detection Demo by Milind Naphade
Video Clip
Detecting semantic concepts using multijects
Outdoor
Rocks
Sky
Snow
Waterbody
Forestry and Greenery
Paused 308 / 9539
Play

Fig. 12. Filtering to view only those shots with a high probability of *rocks*, *water* and *greenery*, and low probability of *sky* and *snow*.

$$\max_{\delta} P_D(\delta) \qquad \text{subject to } P_F(\delta) \le \alpha \tag{7}$$

where α is bound on false-alarm rate. We achieve this by using a likelihood ratio test.

From the Neyman–Pearson ROC curve, we can read the maximum detection rate corresponding to any false alarm rate. The ROC curve gives the user the freedom to vary the threshold depending on requirements as the Neyman–Pearson criterion recognizes the inherent asymmetry in the importance of the two hypotheses. This is especially relevant in the video retrieval scenario where an end user may be a better person to fix thresholds than the system designer.

Fig. 7 shows the ROC curve for the overall performance across all the five multijects. The ROC curve for multiject based detection performance is obtained by using the likelihood ratio test in Equation (8) with the soft decisions at frame level obtained in (6):

$$\frac{P(F_i = 1|\mathcal{X})}{P(F_i = 0|\mathcal{X})} > \tau \qquad 0 \le \tau \le \infty \quad i \in \{1, \cdots, N\} \tag{8}$$

where N is the number of multijects. The curve is obtained by changing the threshold value τ from one extreme [$\tau = 0$ corresponding to the coordinates (1,1) in the graph] to the other [$\tau = \infty$ corresponding to the coordinates (0,0) in the graph]. To obtain overall performance, the performance across all the multijects is averaged. This represents the best possible detection performance using the multijects obtained in Section V. This is then compared against the ROC curve obtained by the likelihood ratio test of (9) using soft decisions at nodes in Layer 1 of the multinet:

$$\frac{P(T_i = 1|F_1, \cdots, F_N)}{P(T_i = 0|F_1, \cdots, F_N)} > \tau$$
$$0 \le \tau \le \infty \quad i \in \{1, \cdots, N\}. \tag{9}$$

Fig. 7 demonstrates significant improvement in detection performance by using the multinet than without using it. Improve-

[4] The authors thank K. Murphy for the BayesNet Toolbox, used to train the Bayesian net.

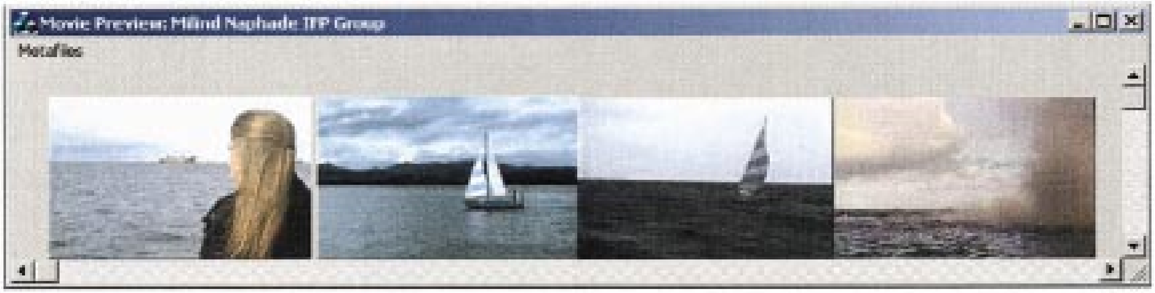

Fig. 13. Retrieving four clips with high probability for keywords *sky* and *water*.

ment in detection (P_d) is more than 20% for a range of thresholds corresponding to small probability of false alarms (P_f).

VIII. Combining Classifiers

Some concepts cannot be modeled using low-level features as they may not offer invariance in these feature spaces. These concepts may however provide invariance in the high-level feature space of multiject-based detectors. For example, it is clear that the presence of one or more of the five site multijects of Section VI boost the chances of an *outdoor* scene. The role of the multinet can then be extended to learn the relation between such a concept and other concepts, which are based on frame-level features. Inference about *outdoor* multiject can be based on the relation it shares with the site multijects. The multinet can thus be trivially extended to infer concepts, which are not represented through frame-level feature-based models.

In some cases we may want to use the relation that a concept shares with other concepts as well as the evidence the multinet receives from a feature-based model of the concept. For example, it is possible to develop a model for the *outdoor* multiject based on low-level global features. In this section we will show how easy it is to extend the multinet for such a task. The fact that the multinet is a Bayesian network is the reason for this simplicity.

As mentioned previously, some concepts exists at region-level, while others at global-level (or frame-level). *Outdoor* and *beach* are two examples of concepts that exist at global level. To distinguish between region-based and global multijects, let us represent the global frame-level multijects by the set of binary random variables G_i, where $i \in \{1, \cdots, N_g\}$ and N_g is the number of global multijects. Defining G_i in (10)

$$G_i = \begin{matrix} 1 & \text{if global concept } i \text{ is present} \\ & \text{in the current frame} \\ 0 & \text{otherwise.} \end{matrix} \tag{10}$$

Let the global feature vector for the frame be $\vec{X}$. The two hypotheses H_0 and H_1 are defined in (11):

$$\begin{aligned} &H_0: \vec{X} \sim P_0\left(\vec{X}\right) \\ &H_1: \vec{X} \sim P_1\left(\vec{X}\right). \end{aligned} \tag{11}$$

$P_0(\vec{X})$ and $P_1(\vec{X})$ denote the probability density functions conditioned on the null hypothesis (concept absent) and the true hypothesis (concept present). These conditional probability density functions are again modeled using a mixture of Gaussian components for the *site* multijects. For the *objects* and *events*, hidden Markov models are used and the feature vectors for all the frames within a shot constitute to the time series modeled by the HMMs.

We propose a modification to the multinet of Fig. 6. First, we extract frame-level or global features. The features include color histogram, color moments, texture and structure. These features are exactly identical to the features described in Section IV except that they are now derived for the whole frame. The model that we develop using these features thus exists directly at the frame-level. Just as described in Section V, we then build a model for the true hypothesis and a model for the null hypothesis. Using these two models and assuming uniform priors on the presence and absence of the global concept we can obtain soft decisions as shown in (12):

$$P\left(G_i = 1 \middle| \vec{X}\right) = \frac{P\left(\vec{X} \middle| G_i = 1\right)}{P\left(\vec{X} \middle| G_i = 1\right) + P\left(\vec{X} \middle| G_i = 0\right)}. \tag{12}$$

We then use the soft decision about the node $outdoor_s$ along with the existing soft decisions of the five other site multijects in Layer 1 of the multinet of Fig. 6. The modified multinet is shown in Fig. 8. The $outdoor_t$ node could have been placed in Layer 1 and connected to all the nodes in Layer 0 just like the other multijects. But it can benefit from the improvement in detection of the other five multijects and is hence defined as the parent of the five nodes in Layer 1. Regional and global data from training-set images is used for training the multinet and data from test-set images is used for testing. Using the soft decisions in (12) and the likelihood ratio test in (13) we obtain the ROC curve for the detection performance of the *outdoor* multiject based on the global multiject model alone:

$$\frac{P\left(outdoor_s = 1 \middle| \vec{X}\right)}{P\left(outdoor_s = 0 \middle| \vec{X}\right)} > \tau \qquad 0 \leq \tau \leq \infty. \tag{13}$$

We then compare this with detection using the soft decisions at node $outdoor_t$ using the likelihood ratio test in (14), shown at the bottom of the next page, where $0 \leq \tau \leq \infty$. The ROC curve of (14) represents detection performance of the *outdoor* multiject using the multinet. The two ROC curves are compared in Figs. 9 and 10.

Fig. 9 reveals that the performance of the Gaussian mixture models based on media features itself is good. The importance of fusion of these heterogeneous soft decisions is evident for very low false alarm rates as shown by a segment of the ROC curve in Fig. 10. The maximum improvement in performance is over 12% for a false alarm probability of 2%.

IX. Multijects and Multinets for Filtering and Semantic Indexing

The block diagram of the system using the multijects and multinet for semantic video indexing is shown in Fig. 11.

We have presented a probabilistic framework of multijects and multinet for semantic video indexing. This framework is designed to handle a large number of multijects. This framework can be used for meaningful filtering of content. For example we may want to view only those clips, which have a high probability of *rocks*, *waterbody*, and *greenery* and a low probability of *sky* and *snow*. Fig. 12 shows a filter playing all video clips with these constraints.

Since the soft decisions are available the user can vary the threshold for each multiject to personalize the filter. For filtering, multijects need to be employed at the client browser if the content is not already preprocessed and indexed. Similarly, multijects for concepts like *explosion*, *gunshots*, etc. can be used to block access to all those video clips on the net, which have graphic depiction of violence. Another example is smart televisions and video recorders, which can scan the available channels, and record all possible video clips with a *beach* or *ball-game*. Semantic indexing can also provide key-word search and bring video clips at par with text-databases. Popular internet search engines can definitely be enhanced if they support key-word based video search. Fig. 13 shows four clips retrieved when searched using the keywords *sky* and *water*.

Since the actual processing is done at the server hosting the video clips or at the search engine through crawlers, the problem of computational cost is not daunting. In fact, once the video clips are automatically annotated using the multijects and multinets, video search reduces to text-search using the keywords. Used in conjunction with the query by example paradigm, this can prove to be a powerful tool for content-based multimedia access.

X. Future Research and Conclusions

We have presented a novel probabilistic framework for semantic video indexing. The framework is based on multijects and multinets. We have used the framework to obtain multiject models for various objects sites and events in audio and video. To discover the interaction between multiject models we have presented a Bayesian multinet and described how it is automatically learnt from examples. Using the multinet to explicitly model the interaction between multijects, we have demonstrated substantial improvement in detection performance and also facilitated detection of concepts, which may not be directly unobserved in the media features. We have also extended the multinet to fuse classifiers and heterogeneous features. Through the framework of multijects and multinets we have proposed and demonstrated an open ended and flexible architecture for semantic video indexing. In addition to the novel probabilistic framework for semantic indexing we have also used an objective quantitative evaluation strategy in the form of ROC curves and have demonstrated the superior detection performance of the proposed scheme using these curves. Future research aims at demonstrating the ability of the multinet to seamlessly integrate multiple media simultaneously and develop multijects for dynamically varying events in video. There is also the need to support dynamically varying relationships amongst semantic concepts. The multinet architecture does not impose any conditions on the multiject architecture except that it provide confidence measures. We can therefore experiment with sophisticated class conditional density functions for modeling multijects. This will lead to an improvement in the baseline performance as well as overall system performance.

References

[1] M. R. Naphade, R. Wang, and T. S. Huang, “Multimodal pattern matching for audio-visual query and retrieval,” *Proc. SPIE, Storage and Retrieval for Media Databases*, Jan. 2001.

[2] D. Zhong and S. F. Chang, “Spatio-temporal video search using the object-based video representation,” in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Santa Barbara, CA, Oct. 1997, pp. 21–24.

[3] Y. Deng and B. S. Manjunath, “Content based search of video using color, texture and motion,” in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Santa Barbara, CA, Oct. 1997, pp. 13–16.

[4] H. Zhang, A. Wang, and Y. Altunbasak, “Content-based video retrieval and compression: A unified solution,” in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Santa Barbara, CA, Oct. 1997, pp. 13–16.

[5] M. M. Yeung and B. Liu, “Efficient matching and clustering of video shots,” in *Proc. IEEE Int. Conf. Image Processing*, Washington, DC, Oct. 1995, pp. 338–341.

[6] M. R. Naphade, M. M. Yeung, and B. L. Yeo, “A novel scheme for fast and efficient video sequence matching using compact signatures,” *Proc. SPIE, IS&T Storage and Retrieval for Multimedia Databases*, vol. 3972, pp. 564–572, Jan. 2000.

[7] B. L. Yeo and B. Liu, “Rapid scene change detection on compressed video,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 533–544, Dec. 1995.

[8] J. Meng, Y. Juan, and S. F. Chang, “Scene change detection in a mpeg compressed video sequence,” in *Proc. IS&T/SPIE Symp.*, vol. 2419, San Jose, CA, Feb. 1995, pp. 1–11.

[9] N. V. Patel and I. K. Sethi, “Video segmentation for video data management,” in *The Handbook of Multimedia Information Management*, W. I. Grosky, R. Jain, and R. Mehrotra, Eds. Upper Saddle River, NJ: Prentice-Hall/PTR, 1997, pp. 139–165.

[10] H. J. Zhang, C. Y. Low, and S. Smoliar, “Video parsing using compressed data,” in *Proc. IS&T/SPIE Conf. Image and Video Processing—II*, San Jose, CA, 1994, pp. 142–149.

[11] M. Naphade, R. Mehrotra, A. M. Ferman, J. Warnick, T. S. Huang, and A. M. Tekalp, “A high performance shot boundary detection algorithm using multiple cues,” in *Proc. Fifth IEEE Int. Conf. Image Processing*, vol. 2, Chicago, IL, Oct. 1998, pp. 884–887.

[12] J. Nam, A. E. Cetin, and A. H. Tewfik, “Speaker identification and video analysis for hierarchical video shot classification,” in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Santa Barbara, CA, Oct. 1997, pp. 550–555.

$$\frac{P(outdoor_t = 1 | outdoor_s,\ rocks_t,\ sky_t,\ snow_t,\ water_t,\ forest_t)}{P(outdoor_t = 0 | outdoor_s,\ rocks_t,\ sky_t,\ snow_t,\ water_t,\ forest_t)} > \tau \qquad (14)$$

[13] M. R. Naphade and T. S. Huang, "Stochastic modeling of soundtrack for efficient segmentation and indexing of video," *Proc. SPIE IS&T Storage and Retrieval for Multimedia Databases*, vol. 3972, pp. 168–176, Jan. 2000.

[14] T. Zhang and C. Kuo, "An integrated approach to multimodal media content analysis," *Proc. SPIE IS&T Storage and Retrieval for Media Databases*, vol. 3972, pp. 506–517, Jan. 2000.

[15] M. Akutsu, A. Hamada, and Y. Tonomura, "Video handling with music and speech detection," *IEEE Multimedia*, vol. 5, no. 3, pp. 17–25, 1998.

[16] M. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to indexing and retrieval in multimedia systems," in *Proc. Fifth IEEE Int. Conf. Image Processing*, vol. 3, Chicago, IL, Oct. 1998, pp. 536–540.

[17] S. F. Chang, W. Chen, and H. Sundaram, "Semantic visual templates—Linking features to semantics," in *Proc. Fifth IEEE Int. Conf. Image Processing*, vol. 3, Chicago, IL, Oct. 1998, pp. 531–535.

[18] R. Qian, N. Hearing, and I. Sezan, "A computational approach to semantic event detection," in *Proc. Computer Vision and Pattern Recognition*, vol. 1, Fort Collins, CO, June 1999, pp. 200–206.

[19] W. Wolf, "Hidden Markov model parsing of video programs," presented at the Int. Conf. Acoustics, Speech, and Signal Processing, 1997.

[20] A. M. Ferman and A. M. Tekalp, "Probabilistic analysis and extraction of video content," in *Proc. IEEE Int. Conf. Image Processing*, Kobe, Japan, Oct. 1999.

[21] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*. Cambridge, MA: MIT Press/McGraw-Hill, 1995.

[22] A. K. Jain and A. Vailaya, "Shape-based retrieval: A case study with trademark image databases," *Pattern Recognit.*, vol. 31, no. 9, pp. 1369–1390, 1998.

[23] A. K. Jain, A. Vailaya, and W. Xiong, "Query by video clip," *Multimedia Syst.*, 1999.

[24] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[25] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann, 1988.

[26] F. V. Jensson, *Introduction to Bayseian Networks*. Cambridge, MA: Springer Verlag, Aug. 1996.

[27] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*. Cambridge, MA: MIT Press, 1998.

[28] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer-Verlag, 1999.

Milind Ramesh Naphade received the B.E. degree in instrumentation and control engineering from the University of Pune, India, in July 1995, ranking first among the university students in his discipline. He received his M.S. degree in electrical engineering from the University of Illinois at Urbana-Champaign (UIUC) in 1998, where he is pursuing the Ph.D. degree in electrical engineering.

He has worked in the applications group with the Center for Development of Advanced Computing (C-DAC), Pune, from July 1994 to July 1996. He was with Kodak Research Laboratories, Eastman Kodak Company, in the summer of 1997 and with Microcomputer Research Laboratories at Intel Corporation in the summer of 1998. Since August 1996, he has been a Research Assistant at the Beckman Institute for Advanced Science and Technology, UIUC. Since August 1999, he has also been awarded a fellowship by the Computational Sciences and Engineering Department of the College of Engineering. His research interests include audio-visual signal processing and analysis for the purpose of multimedia understanding, content-based indexing, and retrieval. He is interested in applying advanced probabilistic pattern recognition and machine learning techniques to model semantics in multimedia data.

Thomas S. Huang (S'61–M'63–SM'76–F'79) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was on the faculty of the Department of Electrical Engineering at MIT from 1963 to 1973, and on the faculty of the School of Electrical Engineering and Director of its Laboratory for Information and Signal Processing, Purdue University, West Lafayette, IN, from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now William L. Everitt Distinguished Professor of electrical and computer engineering, Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology. During his sabbatical leaves he has worked at the MIT Lincoln Laboratory, the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, and the Rheinishes Landes Museum, Bonn, Germany, and held Visiting Professor positions at the Swiss Institutes of Technology, Zürich and Lausanne, University of Hannover, Germany, INRS-Telecommunications, University of Quebec, Montreal, QC, Canada, and the University of Tokyo, Japan. He has served as a Consultant to numerous industrial firms and government agencies both in the U.S. and abroad. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 12 books, and over 400 papers in network theory, digital filtering, image processing, and computer vision. He is a Founding Editor of the *International Journal Computer Vision, Graphics, and Image Processing*; and Editor of the *Springer Series in Information Sciences*, published by Springer Verlag.

Dr. Huang is a Fellow of the International Association of Pattern Recognition and the Optical Society of American, has received a Guggenheim Fellowship, an A. V. Humboldt Foundation Senior U.S. Scientist Award, and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Signal Processing Society's Technical Achievement Award in 1987 and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. Also in 2000, he received the Honda Lifetime Achievement Award for "contributions to motion analysis."