A Factor Graph Framework for Semantic Video Indexing

Milind Ramesh Naphade, Igor V. Kozintsev, and Thomas S. Huang

***Abstract*—Video query by semantic keywords is one of the most challenging research issues in video data management. To go beyond low-level similarity and access video data content by semantics, we need to bridge the gap between the low-level representation and high-level semantics. This is a difficult multimedia understanding problem. We formulate this problem as a probabilistic pattern-recognition problem for modeling semantics in terms of concepts and context. 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 appear in context. To model this interaction explicitly, we propose a network of multijects (*multinet*). To model the multinet computationally, we propose a factor graph framework which can enforce spatio-temporal constraints. Using probabilistic models for multijects, *rocks, sky, snow, water-body,* and *forestry/greenery*, and using a factor graph as the multinet, we demonstrate the application of this framework to semantic video indexing. We demonstrate how detection performance can be significantly improved using the multinet to take inter-conceptual relationships into account. Our experiments using a large video database consisting of clips from several movies and based on a set of five semantic concepts reveal a significant improvement in detection performance by over 22%. We also show how the multinet is extended to take temporal correlation into account. By constructing a dynamic multinet, we show that the detection performance is further enhanced by as much as 12%. With this framework, we show how keyword-based query and semantic filtering is possible for a predetermined set of concepts.**

***Index Terms*—Factor graphs, hidden Markov models, likelihood ratio test, multimedia understanding, probabilistic graphical networks, probability propagation, query by example, query by keywords, ROC curves, semantic video indexing, sum-product algorithm.**

I. Introduction

THE availability of digital video content has increased tremendously in recent years. Rapid advances in the technology for media capture, storage, and transmission, and the dwindling prices of these devices, has contributed to an amazing growth in the amount of multimedia content that is generated. While content generation and dissemination grows explosively, there are very few tools to filter, classify, search, and retrieve it efficiently. Lack of tools for efficient content-based access and multimedia data mining threatens to render most of this data useless. Current techniques in content-based retrieval for image sequences support the paradigm of query by example using similarity in low-level media features [1]–[6]. 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 database clips according to a heuristic measure of similarity between the query and the target.

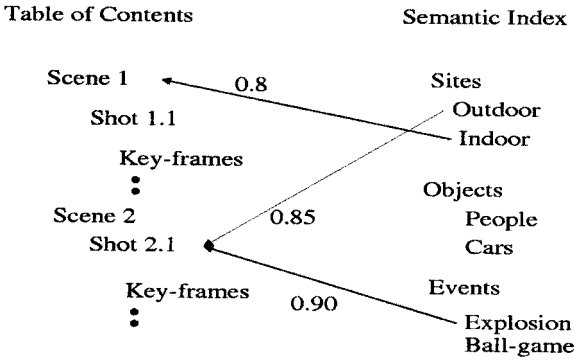


Fig. 1. Organizing a video with a ToC and SI). The ToC gives a top-down break-up in terms of scenes, shots and key-frames. The SI lists key concepts occurring in the video. The links indicate the exact location of these concepts and the confidence measure.

Although effective for browsing and low-level search, there are some basic problems, with this paradigm. The first is that low-level similarity may not match with the user's perception of similarity. This is often caused by the user having high-level semantics in mind, which the system cannot support. Secondly, it is not realistic to assume that a person may have access to a clip, which can represent, what the person wishes to find. In order to be able to analyze content semantically, it is also essential to fuse information from multiple modalities, especially the image sequence and audio streams. Several existing systems fail to use this multimodality. To address these problems, we need a semantic indexing, filtering, and retrieval scheme, which can map low-level multimodal features to high-level semantics.

One way of organizing a video for efficient browsing and searching is shown in Fig. 1. On one hand, there is a systematic top-down breakdown of the video into scenes, shots and key-frames. On the other hand there is a semantic index (SI), which lists the key semantic concepts occurring in the shots. The links connecting entries in the SI to shots/scenes in the Table of Contents (ToC) also indicate a measure of confidence about the occurrence of the particular concepts in the video.

Manuscript received June 23, 2000; revised October 19, 2001. This paper was recommended by Associate Editor H.-J. Zhang.

M. R. Naphade is with IBM T.J. Watson Research Center, Hawthorne, NY 10532 USA and also with the Department of Electrical and Computer Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign, IL 61801 USA (e-mail: milind@ifp.uiuc.edu).

I. V. Kozintsev and T. S. Huang are with the Department of Electrical and Computer Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign, IL 61801 USA (e-mail: igor@ifp.uiuc.edu; huang@ifp.uiuc.edu.)

Publisher Item Identifier S 1051-8215(02)01127-8.

Automatic techniques for generating the ToC exist. The first step in generating the ToC is the segmentation of the video track into smaller units. Shot boundary detection can be performed in compressed domain [7]–[9], as well as uncompressed domain [10]. Shots can be grouped based on continuity, temporal proximity, and similarity to form scenes [5]. Most systems supporting query by example [2]–[6] can be used to group shots and enhance the ability to browse. The user may browse a video and then provide one of the clips in the ToC structure as an example to drive the retrieval systems mentioned earlier. Chang *et al.* [2] allow the user can provide a sketch of a dominant object along with its color shape and motion trajectory. Key-frames can be extracted from shots to help efficient browsing. The ToC is thus useful in efficient browsing. The need for a semantic index is felt to facilitate search using keywords or key concepts. For a system to fetch clips of an aeroplane, the system must be able to capture the semantic concept of an aeroplane in terms of a model. Similarly, to support a query of the *explosion on a beach* kind, the system must understand how the concepts *explosion* and *beach* are represented. This is a difficult problem. The difficulty lies in the gap that exists between low-level media features and high-level semantics. Query using keywords representing semantic concepts has motivated recent research in semantic video indexing [11]–[14]. Recent attempts to introduce semantics in the structuring of videos includes [15]–[17]. We present novel ideas in semantic video indexing by learning probabilistic multimedia representations of semantic events like *explosion* and sites like *waterfall* [11]. Chang *et al.* introduce the notion of semantic visual templates [12]. Wolf *et al.* use hidden Markov models to parse video [15]. Ferman *et al.* attempt to model semantic structures like *dialogues* in video [16].

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 pattern-recognition techniques to develop models representing semantic concepts and show how these models can be used for filtering of semantic concepts and retrieval using keywords. 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 can be 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 the performance of the scheme. 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*. Using these concepts for our experiments, we demonstrate how filtering and keyword-based retrieval can be performed on multimedia databases.

The paper is organized as follows. In Section II, we present a probabilistic framework of multijects and multinets to map low-level features to semantics in terms of concepts and context. In Section III, we discuss the preprocessing steps including feature extraction, representation, segmentation and tracking. We also discuss the details of the database used in subsequent experiments. In Section IV, we present the experimental results of detection for a set of five semantic concepts with and without using the multinet to model spatial as well as spatio-temporal context. We show how the use of the multinet enhances detection performance in Section IV. We also discuss the application of this framework to filter videos semantically in Section IV. Directions for future research and conclusions are presented in Section V.

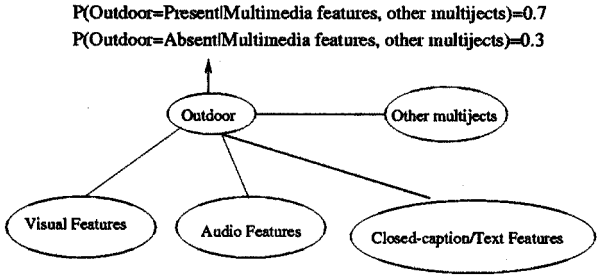


Fig. 2. A multiject for the semantic concept *outdoor*. The media support for the label *outdoor* is in the form of audio-visual features. In addition to this, there may be support from other multijects representing semantic concepts like *sky*.

II. Probabilistic Framework of Multijects and Multinets

A. Probabilistic Multimedia Objects (Multijects)

Users of video 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 using low-level features only. To support such queries, we define a *multiject*. A multiject [11] 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.

B. The Multinet: A Network of Multijects

Semantic concepts are related to each other. One of the main contributions of this paper is a computational framework in the form of a probabilistic graphical network 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 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 the multijects and model their interaction or context, we therefore propose a network of multijects, which we call a

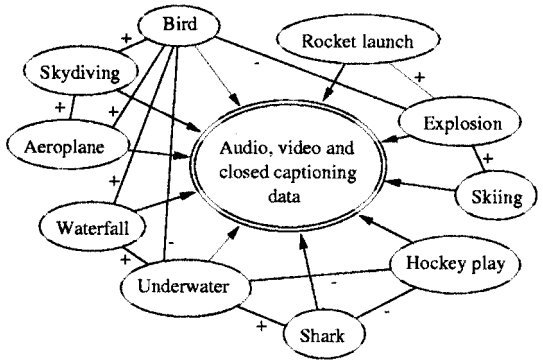


Fig. 3. Conceptual figure of a multinet. The multinet captures the interaction between various semantic concepts. The edges between multijects denote interaction and the signs on the edges denote the nature of interaction.

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 taking into account the relationship between various multijects, we can enforce spatio-temporal constraints to do the following.

1) Enhance Detection: Detection of multijects can be enhanced by correcting the soft decisions based on the constraints enforced by the context.

2) Support Inference: While some multijects may be easier to detect, others 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 relation that these multijects share 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*, and *boat*. Based on this detection of *beach*, we can then claim that the scene is an *outdoor* scene.

3) 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.

The multinet is thus a mechanism for imposing spatio-temporal constraints governing the joint existence of semantic concepts with spatio-temporal support.

C. Estimating Multiject Models

The multijects link the low-level features to high-level labels through a probabilistic structure. Depending on the support, a multiject enjoys in space and time, the structure in which the features are probabilistically encoded varies. In general, the multiject might either enjoy only spatial support statically within a frame or enjoy spatio-temporal support in an image sequence or audio frame sequence. We build our multiject models using the Bayes decision theory [18], [19]. Let each observation (image/audio frame) be represented in terms of a feature vector $\vec{X}$. We characterize these features through their statistical properties. We assume that the features are drawn from probability distribution functions under all possible mutually exclusive hypotheses. Under each hypothesis, we define a class-conditional density function for the features and a prior on the hypothesis. We assume that, while using the Bayes decision theory to choose among the possible hypotheses, these class-conditional density functions are known to us through estimation. In the simplest form, we model a semantic concept as a binary random variable and define the two hypotheses H_0 and H_1 as

$$\begin{aligned} H_0&:\vec{X} \sim P_0(\vec{X}) \\ H_1&:\vec{X} \sim P_1(\vec{X}) \end{aligned} \tag{1}$$

where $P_0(\vec{X})$ and $P_1(\vec{X})$ denote the class-conditional probability density functions of the feature vectors conditioned on the null hypothesis (concept absent) and the true hypothesis (concept present) respectively. In case of sites (or static patterns), these class-conditional density functions of the feature vector under the true and null hypotheses are modeled as mixture of multidimensional gaussians (Gaussian mixture models or GMMs). The temporal flow is not taken into consideration. In case of events and objects with spatio-temporal support, we use hidden Markov models (HMM) with continuous multidimensional 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. In the case of temporal support, $\vec{X}$ is assumed to represent the time series of the feature vectors within a single video shot. Assuming that the class conditional density functions are known to us and that the cost of making a decision α_i, when the true class is ω_j is λ_{ij} and is defined in (3)

$$\lambda_{ij} = \begin{matrix} 0, & i = j \\ 1, & \neq j \end{matrix}. \tag{2}$$

We can use the Bayes decision rule to choose hypothesis H_1 over H_0 for a new image/image sequence represented by features $\vec{X}'$ if

$$\frac{P_1(\vec{X}')}{P_0(\vec{X}')} > \frac{P(\omega_0)}{P(\omega_1)}. \tag{3}$$

Otherwise, we choose hypothesis H_0. The test in (3) is known as the likelihood ratio test [18].

As stated earlier, the use of (3) demands the knowledge of the class conditional density functions. To evaluate these functions using the maximum likelihood parameter estimation technique, the EM algorithm [20], [21] is used in both cases to estimate the means, covariance matrices, mixing proportions (GMM and HMMS) and transition matrix (HMM).

In this paper, we present results of the detection using (3) in Section IV for the following regional site multijects which use visual features alone: *sky, water, forest, rocks*, and *snow*. While these five site multijects are used in our experiments in the remainder of this paper, we have developed models for several other multijects. Some of them are based on audio features, e.g., *human-speech, music* [22], and *helicopter*. Others are based on image sequence features e.g., *outdoor* [23], *beach*, etc. Some others are based on audio and video features e.g., *explosion, waterfall* [11]. Through these examples, we have demonstrated that this framework is scalable. As long as the concepts belonging to the three types—objects, sites, and events—offer some invariance in one or more features and there is a large training set for estimating class-conditional density functions, we can model the multiject and, thus, estimate the probabilistic mapping from

low-level features to high-level semantics. While the construction of a large labeled training set can be a tedious procedure, we have presented some preliminary results in alleviating the burden of labeling large data-sets [24].

An important aspect of modeling semantics is the interaction between semantic concepts, that forms the context. Humans use their knowledge to enforce context. In Section II-D, we present an elegant computational framework to model context in terms of co-occurrence.

D. Factor Graphs

To model the interaction between multijects in a multinet, we propose using a novel *factor graph* [25]–[27] framework. A factor graph is a bipartite graph that expresses how a *global* function of many variables factors into a product of *local* functions [26], [27]. Factor graphs subsume many other graphical models including Bayesian networks and Markov random fields. Many problems in signal processing and learning are formulated as minimizing or maximizing a global function $f(\mathbf{x})$ marginalized for a subset of its arguments. The algorithm which allows us to perform this efficiently, though in most cases only approximately, is called the **sum-product algorithm**. Based on a rule, the *sum-product algorithm* [27] is used in factor graphs to compute various marginal functions by distributed message-passing. Depending on the structure of the global function represented by the factor graph, the sum–product algorithm can lead to exact or approximate computation of the marginal functions. Many algorithms in various engineering and scientific areas turn out to be examples of the sum-product algorithm. Famous examples include the BCJR algorithm [28], the forward–backward algorithm [21], Pearl's belief propagation and belief revision algorithm [29] operating in a Bayesian network.

Factor graphs were initially successfully applied in the area of channel-error correction coding [30], [31] and specifically, iterative decoding [32], [33]. Turbo decoding and other iterative decoding techniques have, in the last few years, proven to be landmark developments in coding theory. Before explaining how factor graphs can be used to model global functions we introduce some necessary notation. Most of the notation here follows Kschischang *et al.* [27]. Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ be a set of variables. Consider a function $f(\mathbf{x})$, with factors as follows:

$$f(\mathbf{x}) = \prod_{i=1}^{m} f_i\left(\mathbf{x}^{(i)}\right) \tag{4}$$

where $\mathbf{x}^{(i)}$ is the set of variables, which are the arguments of the function f_i. A factor graph for f is defined as the bipartite graph with two vertex classes V_f and V_v of sizes m and n respectively, such that the ith node in V_f is connected to the jth node in V_v if and only if x_j is an argument of function f_i. Fig. 4 shows a simple factor graph representation of $f(x_1, x_2, x_3) = f_1(x_1, x_2) * f_2(x_2, x_3)$ with function nodes f_1, f_2 and variable nodes x_1, x_2, x_3.

E. The Sum–Product Algorithm

The sum–product algorithm works by computing messages at the nodes using a simple rule and then passing the messages between nodes according to a selected schedule. For a discussion on schedules, see Kschischang *et al.* [27]. A message from a function node to a variable node is the product of all messages incoming to the function node with the function itself, summarized for the variable associated with the variable node. A message from a variable node to a function node is simply the product of all messages incoming to the variable node from other functions connected to it.

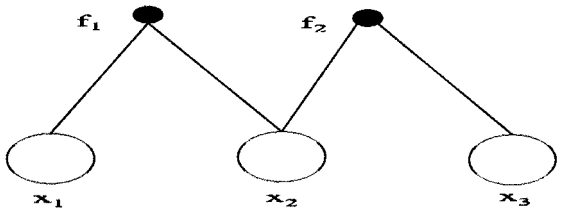


Fig. 4. Simple factor graph with a factorization of $f(x_1, x_2, x_3)$ as $f_1(x_1, x_2) * f_2(x_2, x_3)$.

Consider a message on the edge connecting function node f to variable node v. Let $\mathrm{mesg}_{v \to f}$ denote the message sent along the edge $\{f, v\}$ from variable node v to function node f. Also, let $\mathrm{mesg}_{f \to v}$ denote the message sent along the edge $\{f, v\}$ from function node f to function node v. Further, let $n(x)$ denote the set of all the neighbors of node x and let $\downarrow$ indicate the summary operator. A summary operator summarizes a function for a particular set of variables. For example consider a function $f(x_1, x_2, x_3)$, then a possible summary operator could be the summation operator in (5)

$$f(x_1, x_2, x_3) \downarrow x_1 = \sum_{x_2, x_3} f(x_1, x_2, x_3). \tag{5}$$

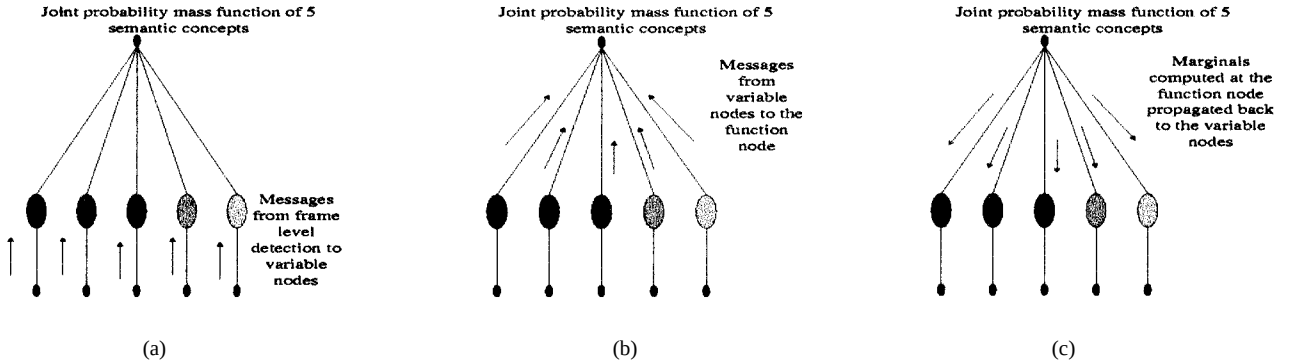
With this notation, the message computations performed by the sum–product algorithm can be expressed as follows in (6) and (7):

$$\mathrm{mesg}_{v \to f}(v) = \prod_{j \in n(v)/\{f\}} \mathrm{mesg}_{j \to v}(v) \tag{6}$$

$$\mathrm{mesg}_{f \to v}(v) = \left(f(v_{n(f)}) \prod_{k \in n(f)/\{v\}} \mathrm{mesg}_{k \to f}(k) \right) \downarrow v. \tag{7}$$

Probability propagation in Bayesian nets [29] is equivalent to the application of the sum–product algorithm to the corresponding factor graph. If the factor graph is a tree, exact inference is possible using a single set of forward and backward passage of messages. For all other cases, inference is approximate and the message passing is iterative [27], leading to loopy probability propagation. This has a direct bearing on our problem because relations between semantic concepts are complicated and, in general, contain numerous cycles (e.g., see Fig. 3).

The single most significant outcome of using a factor graph framework for a multinet is that the interaction between semantic concepts need not be modeled as a causal entity. The next most significant outcome is that loops and cycles can be supported. It must be noted, though, that when the factor graph is not a tree, the marginals computed are approximate.



(a) (b) (c)

Fig. 5. A multinet: accounting for concept dependencies using a function equivalent to the joint mass function of five concepts. (a) Passing the messages $P(F_i = 0 \mid \mathcal{X})$ and $P(F_i = 1 \mid \mathcal{X})$ to the variable nodes. (b) Propagating the messages received in (a) to the function node. (c) Propagating the messages from the function node back to the variable nodes after appropriate marginalization.

F. Factor Graph Multinet

We now describe a frame-level factor graph to model the probabilistic relations between various frame-level semantic features F_i defined in

$$F_i = \begin{matrix} 1, & \text{if concept } i \text{ is present in the current frame} \\ 0, & \text{otherwise.} \end{matrix} \tag{8}$$

To capture the co-occurrence relationship between the semantic concepts at the frame level, we define a function node which is connected to the variable nodes representing the concepts, as shown in Fig. 5(a). This function node represents the joint-probability mass function of the five semantic concepts represented at frame level by the binary random variables $F_i, i \in \{1, \ldots, N\}$ i.e., $P(F_1, F_2, F_3, \ldots, F_N)$. The joint function over all the random variables in the factor graph is then given by (9)

$$\prod_{i=1}^{i=N} P(\mathcal{X} \mid F_i) P(F_1, \ldots, F_N). \tag{9}$$

Each entry in the joint mass function table tells us about the numerical viability of the configuration of the N random variables. For example, if there are only two concepts, *outdoor* (F_1) and *helicopter* (F_2), the entry corresponding to the configuration $F_1 = 1, F_2 = 0$ tells us how likely it is to be outdoor, without seeing or hearing a helicopter given our model of the world (context), while the entry corresponding to the configuration $F_1 = 0, F_2 = 1$ tells us how likely it is to hear or see a helicopter in an indoor scene. Clearly, one would imagine that it is more likely to see a helicopter while in an outdoor scene $(F_1 = 1, F_2 = 1)$ than in an indoor scene $(F_1 = 0, F_2 = 1)$. Another observation is that if we are presented with very strong evidence of having heard or seen a helicopter, this should boost out belief of being in an outdoor scene. It is through intuitive interactions like the ones mentioned in this example that the multinet fuses context with evidence.

The function nodes below the five variable nodes in Fig. 5 denote the frame-level soft decisions for the binary random variables F_i given the image features $\mathcal{X}$,i.e., $P(F_i = 0 \mid \mathcal{X})$ and $P(F_i = 1 \mid \mathcal{X})$. These are then propagated to the function node. At the function node, the messages are multiplied by the function, which is estimated from the co-occurrence of the concepts in the training set. The function node then sends back messages summarized for each variable. This modifies the soft decisions at the variable nodes according to the high-level relationship between the five concepts. The probability mass function at the function node in Fig. 5 is exponential in the number of concepts (N) and computational cost may increase quickly. To alleviate this, we can enforce a factorization of the function in Fig. 6 as a product of a set of local functions where each local function accounts for the co-occurrence of two variables only. Fig. 6 shows one iteration of message passing in the multinet with a factored joint mass function.

Each function in Fig. 6 represents the joint probability mass of those two variables that are its arguments (and there are C_2^N such functions), thus reducing the complexity. The joint function over all the random variables in the factor graph is now given by

$$\prod_{i=1}^{i=N} P(\mathcal{X} \mid F_i) \prod_{j,k} P(F_k, F_j), \quad \text{where } j < k \quad \text{and} \quad j, k \in \{1, \ldots, N\}. \tag{10}$$

The factor graph is no longer a tree and exact inference becomes hard as the number of loops grows. We then apply iterative message passing based on the sum-product algorithm to overcome this. Each iteration involves a set of messages passed from variable nodes to function nodes and a set of messages passed from the function nodes back to the variable nodes. Approximate marginals are obtained after a few iterations of message passing. The most significant achievement of the factorized multinet is that it makes the computational model of context scalable. Estimating entries at each local function is highly efficient computationally and estimating all the entries for the C_2^N functions is quadratic in the number of concepts. This is a significant improvement computationally, as compared to the estimation for the global function, which was exponential in the number of concepts. The second most significant achievement is the ability to model causal as well as noncausal interactions. It is this ability that makes factor graphs an elegant framework to implement the multinet as against causal probabilistic graphical networks like the Bayesian networks [29], [34].

G. Dynamic Multinets: Extending the Dependence Temporally

In addition to the inter-conceptual intra-frame dependencies, we can also model the inter-frame, intra-conceptual dependencies. Since processing is temporally localized to frames within

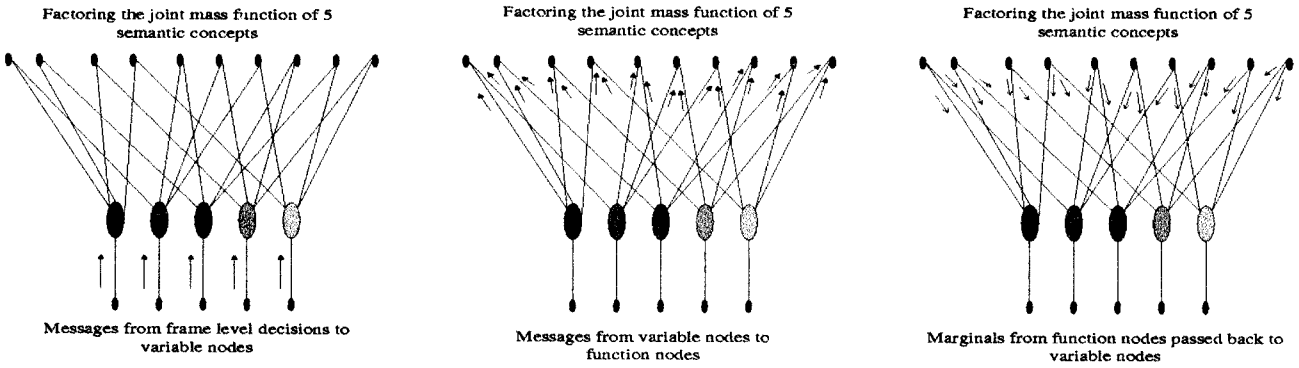


Fig. 6. Replacing the unfactored function in Fig. 5 by a product of ten local functions. Each local function now accounts for the co-occurrence of only two variables. (a) Passing the messages computed in (15) to the variable nodes. (b) Propagating the messages received in (a) to the function node. (c) Propagating the messages from the function node back to the variable nodes after appropriate marginalization.

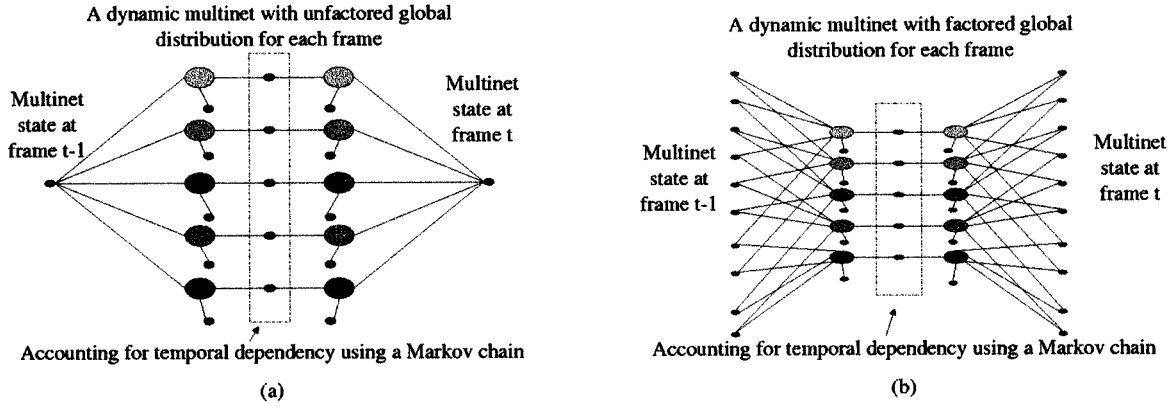


Fig. 7. (a) Replicating the multinet in Fig. 5 for each frame in a shot and introducing temporal dependencies between the binary random variable representing identical concepts at frame level in consecutive frames. The function node that appears below each variable node represents the message computed in (15). (b) Repeating this for Fig. 6.

a shot, there is low probability of a concept appearing in a frame and disappearing in the next frame. Modeling this temporal dependency for each concept can lead to smoothing of the soft decisions within each shot. These dependencies can be modeled by extending the multinets in Figs. 5 and 6, as shown in Fig. 7.

The multinets in Figs. 5 and 6 represent a single slice or video frame. We replicate the slice of factor graphs in Figs. 5 or 6 as many times as the number of frames within a single video shot. Between the nodes in consecutive slices, representing identical concepts, we now introduce a function which captures the dynamics of this concept across frames. For a concept F_i, let $t-1$ and t represent consecutive frames. Then the function represents the transition matrix A_i as

$$A_i = \begin{bmatrix} P(F_i^t = 0 \,|\, F_i^{t-1} = 0) & P(F_i^t = 0 \,|\, F_i^{t-1} = 1) \\ P(F_i^t = 1 \,|\, F_i^{t-1} = 1) & P(F_i^t = 1 \,|\, F_i^{t-1} = 1) \end{bmatrix}.$$

Fig. 7(a) and (b) show two consecutive time slices and extend the models in Figs. 5 and 6, respectively. The horizontal links in Fig. 7(a) and (b) connect the variable node for each concept in a time slice to the corresponding variable node in the next time slice through a function modeling the transition probability. This framework now becomes a dynamic probabilistic network.

For inference, messages are iteratively passed locally within each slice. This is followed by message passing across the time slices in the forward direction and then in the backward direction. Accounting for temporal dependencies thus leads to temporal smoothing of the soft decisions within each shot.

III. Experimental Setup, Preprocessing, and Feature Extraction

A. 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. 8 shows a random 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[1]. The MPEG streams of data are decompressed to perform shot-boundary detection, spatio-temporal video-region segmentation 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. In effect we are using a

[1]We have made an attempt to represent several genres of movies in the database making it a collection of videos with significant variability. Unfortunately no standard databases are available to us for the purpose of benchmarking. We hope that with increasing interest of the multimedia retrieval community in semantic indexing, a benchmark database will emerge.

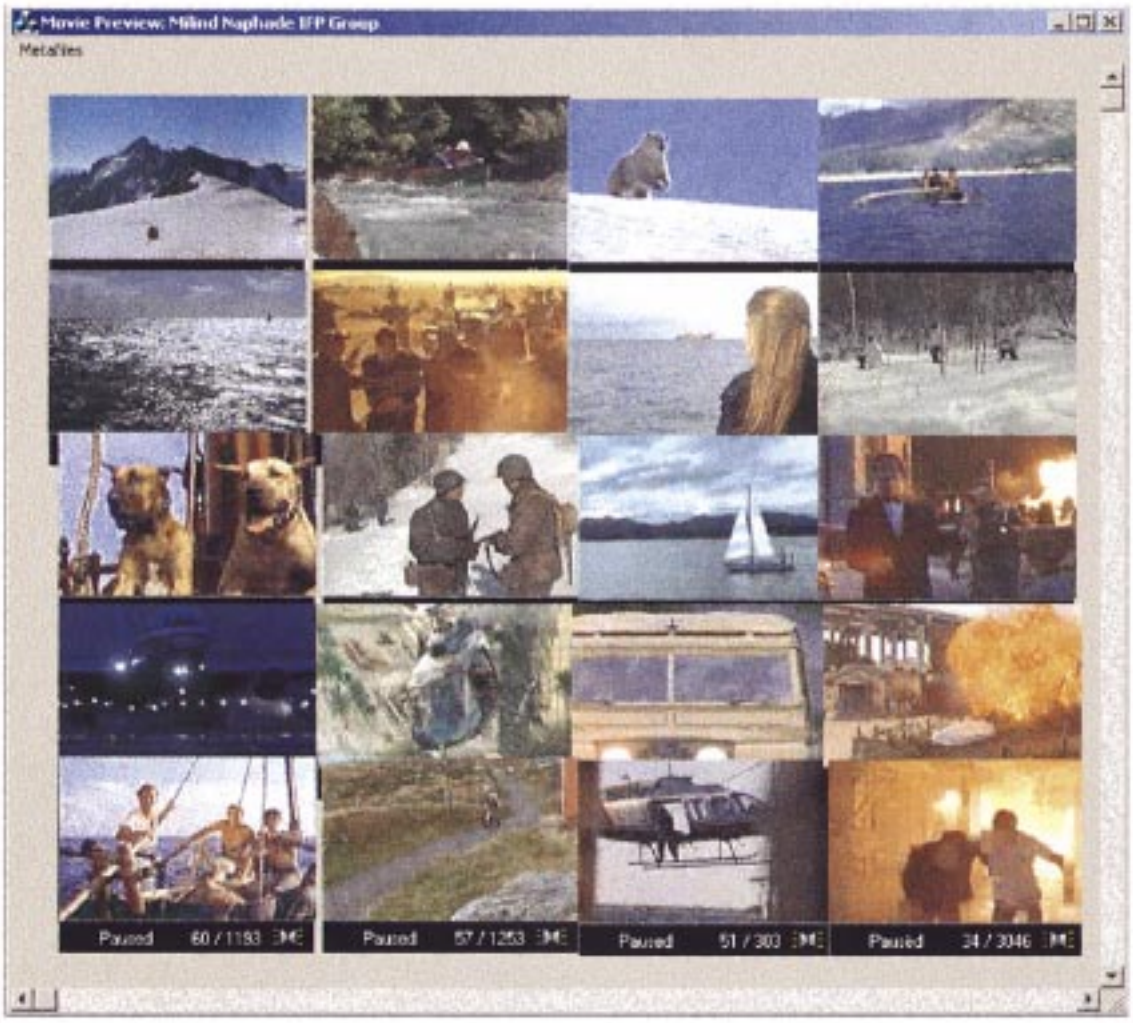


Fig. 8. Random collection of shots from some of the movies in the database.

training set of 18 000 frames and a test set of 94 000 frames. Each frame in the video is of the size 176×112 pixels. For each concept, the model for the true hypothesis has five components in the Gaussian mixture. For each concept, the model for the null hypothesis has ten components in the gaussian mixture. The reason for having more components for the model for the null hypothesis is that the null hypothesis is expected to cover a lot more variations than the true hypothesis. 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.

Some multijects exist at the region level (*face*), while others exist at the global or frame level (*outdoor*). To build multiject models, we need to extract features at regional and global level from the visual stream and features from the audio stream as well.

B. Preprocessing and Feature Extraction

The video clips are segmented into shots using the algorithm by Naphade *et al.* [10]. We then use the spatio-temporal segmentation in [2] applied separately to each shot to obtain regions homogeneous in color and motion. Depending on the genre 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 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. Fig. 9 shows a video frame and its segmented version with six dominant segments. These segments 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.

Fig. 9. Spatio-temporal segmentation applied to each shot. (a) Frame from the sequence. (b) Segmented version of the sequence.

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

1) Color: A normalized, linearized[3] 3-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.

2) Texture: Texture is a spatial property. A 2-D 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, we denote

$$p(i,j,d,\theta) = \frac{P(i,j,d,\theta)}{N(d,\theta)} \tag{11}$$

where $P(\cdot)$ is the GLCM for the displacement vector d and orientation θ and $N(\cdot)$ is the normalizing factor to make the left-hand side of (11) 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 1 unit from the current pixel respectively $(d = 1)$. For each of the four matrices (corresponding to a fixed d and θ), six statistical features of the GLCMs are computed. The features are Contrast, Energy, Entropy, Homogeneity, Correlation, and Inverse Difference Moment [35].

3) 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 [36]. The edge direction histogram is supposed to be a robust representation of shape [37].

4) Shape: Moment invariants as in Dudani *et al.* [38] are used to describe shape of each region. For a binary image mask the central moments are given by

$$\mu_{pq} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^p (y_i - \bar{y})^q \tag{12}$$

where x, y are the image coordinates, $\bar{x}$ and $\bar{y}$ are the mean values of the x and y coordinates, respectively, and the order of the central moment μ_{pq} is $p + q$.

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

6) Color Moments: The first-order moments and the second-order central moments are computed for each of the three channels H, S, and V.

[2]Our aim is to work with a set of reasonable features. There is no claim to the optimality of this set of features and it is definitely endorsed that better features will lead to better performance.

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

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/subset of features can also be obtained at the global level without segmentation and also on difference frames obtained using successive consecutive frames [11].

Semantic concepts like *explosion, helicopter-flying, man-talking*, etc. are heavily dependent on the audio features. In this paper, we will deal with multijects which only use visual features. Details about audio feature extraction and audio models for concepts like *explosion* [11], *music* [22], etc. are not presented here. Some semantic concepts enjoy local or regional support in the image sequence. Examples include sites like *sky* or *water-body*. Some others enjoy global support (over the entire frame). Examples include *outdoor* or *beach*. If a concept enjoys regional support, the probability that a concept occurs in a particular frame given the features for all the regions in the frame is a function of the probabilities, with which it occurs in these regions. To obtain a single frame-level/global measure of confidence, we therefore need to integrate region-level soft decisions (confidence measures). The multinet exists at the frame level and uses these frame-level soft decisions to model context.

C. Integrating Regional Multijects to the Frame Level

A static 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 into account the unavoidable imperfections in segmentation. We tune the segmentation algorithm to obtain coarse, large regions. This can lead to the existence of multiple semantic concepts in a single segment or region. We address this problem by checking each region or segment 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. For the binary classification of each concept in each region, we define binary random variables R_{ij} here

$$R_{ij} = \begin{cases} 1, & \text{if concept } i \text{ is present in region } j \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Using the Bayes' rule, we then obtain (14), shown at the bottom of the page, where X_j denotes the feature vector for region j. The multijects used here are region-level semantic-detectors. To integrate them at the frame-level we define frame-level semantic features $F_i, i \in \{1, \ldots, N\}$ defined in (8). To fuse the region-level concepts we use the OR operator. Let the number of regions in the frame be M. Using the compact notation $\mathcal{X} = \{\vec{X_1}, \ldots, \vec{X_M}\}$, the OR operation is as defined as

$$P(F_i = 0 \,|\, \mathcal{X}) = \prod_{j=1}^{M} P(R_{ij} = 0 \,|\, \vec{X_j})$$
$$P(F_i = 1 \,|\, \mathcal{X}) = 1 - P(F_i = 0 \,|\, \mathcal{X}). \tag{15}$$

If we extract features at the frame-level (globally), then the multijects based on such features exist at the frame level and there is no need for fusion across regions.

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	Accurate Detection	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%

IV. RESULTS

The detection performance of the five *site* multijects over the test-set is given in Table I. The results in Table I are based on a maximum likelihood binary classification strategy using the GMMs for the true and null hypotheses for each multiject.

A. Using the Factor Graph Multinet for Enhancing Detection

We use the soft decisions of the multijects in the frames from the training set to train the multinet. To evaluate the performance of the system over the frames in the test-set, we propose to compare the detection performance over the test-set using the receiver operating characteristics (ROC) curves. An ROC curve is one of the most explicit methods of performance evaluation for binary classification. An ROC curve is a parametric plot of the probability of detection plotted against the probability of false alarms obtained at different values of the parameter (the threshold in our case). A false alarm occurs, when a concept is detected by the scheme, while it is not present. Detection is defined as detecting a concept when it is is actually present. Any point on the ROC curve thus corresponds to the best possible detection performance using the likelihood ratio test [18] subject to the particular false alarm rate using that detection scheme. Operating at various probabilities of false alarms, one clearly wants to attain the highest probability of detection possible.

Fig. 10 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 (16) with the soft decisions at the frame level obtained in (15)

$$\frac{P(\mathcal{X} \mid F_i = 1)}{P(\mathcal{X} \mid F_i = 0)} > \tau 0 \leq \tau \leq \infty, \quad i \in \{1, \ldots, N\} \qquad (16)$$

where N is the number of multijects. The different points on the ROC curve are 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.). We evaluate the probability of detection and false alarm at 2000 threshold values to obtain the curves. 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 II-C. This is then compared against the ROC curve obtained by the likelihood ratio test using soft decisions after a forward and backward pass of messages in the multinet of Fig. 5(c). The third ROC curve is obtained by using the soft decisions after several iterations of loopy probability propagation through message passing in the multinet of Fig. 6.

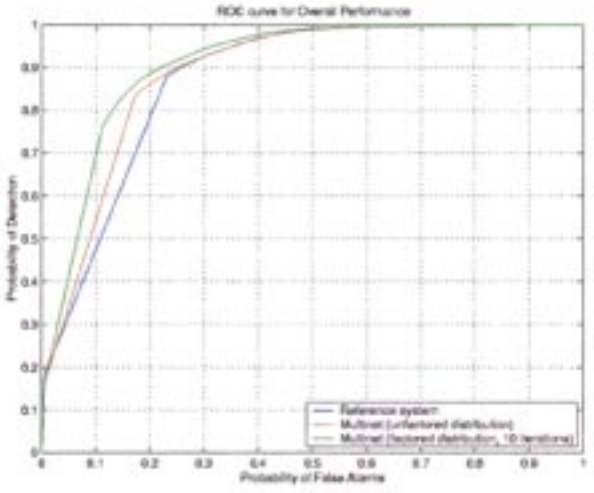


Fig. 10. ROC curves for overall performance using the multijects for isolated detection, the factor graph in Fig. 5, and the graph in Fig. 6.

Fig. 10 demonstrates excellent improvement in detection performance by using the multinets in Fig. 5 over the isolated detection using frame level multiject-based features of (15). Interestingly, detection based on the factorized function (Fig. 6) performs better than the the one based on the unfactorized function (Fig. 5). This may suggest that the factorized function of Fig. 6 is a better representative for the concept dependencies than the one shown in Fig. 5 due to the fact that the factorized function is estimated more reliably (it has less coefficients to estimate). There is also the possibility that local interactions within subsets of concepts are stronger and are better characterized than global interactions. Improvement in detection (P_d) is more than 22% for a range of thresholds corresponding to small probability of false alarms (P_f). To compare the joint detection performance of the system with and without the multinet, we also plot probability of error curves. In order to accommodate all possible 2^N hypotheses (corresponding to every possible configuration of the N binary random variables), we view each configuration as a hypothesis and then use the one-zero cost function in (3). Fig. 11 shows that. irrespective of our choice of threshold, classification error is least for the multinet with factorized joint mass.

$$P(R_{ij} = 1 \mid \vec{X}_j) = \frac{P(\vec{X}_j \mid R_{ij} = 1)P(R_{ij} = 1)}{P(\vec{X}_j \mid R_{ij} = 1)P(R_{ij} = 1) + P(\vec{X}_j \mid R_{ij} = 0)P(R_{ij} = 0)} \qquad (14)$$

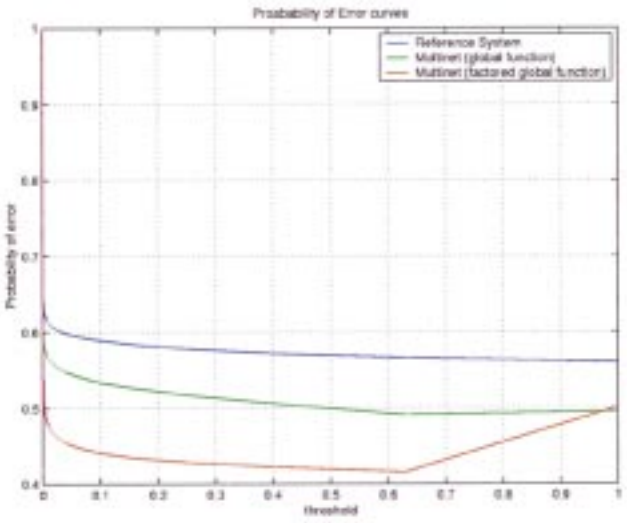


Fig. 11. Probability of error curves for the baseline system, the multinet with the unfactored global function and the multinet with the factorized global function. The multinet with factorized form results in the lowest error for any threshold value.

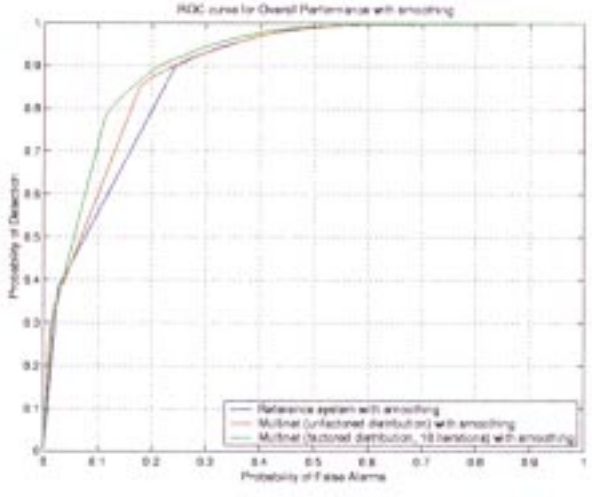


Fig. 12. ROC curves for overall performance using temporal smoothing. The curves correspond to performance using multijects for isolated detection, the factor graph in Fig. 7(a) and the graph in Fig. 7(b).

B. Using the Dynamic Multinet to Model Spatio-Temporal Context

The baseline performance is now obtained by using the frame-level multiject features obtained in (15) followed by temporal smoothing using the forward backward propagation within frames in a shot. This is then compared to the performance obtained by using the dynamic multinets in Fig. 7(a) and (b). Once again, the performance of the multinet with factored global distribution and temporal smoothing is superior to the other configurations. Also, the performance of the multinet with unfactored global distribution and temporal smoothing is better than the baseline. To compare the performance with and without temporal smoothing, we compare the three configurations individually. The comparison can be seen in Figs. 13–15.

From Figs. 13–15, the benefit of temporal smoothing is obvious. For each configuration, there is further improvement in detection performance by using the multinet to model the dependencies between intra-frame concepts (Fig. 7) and inter-frame

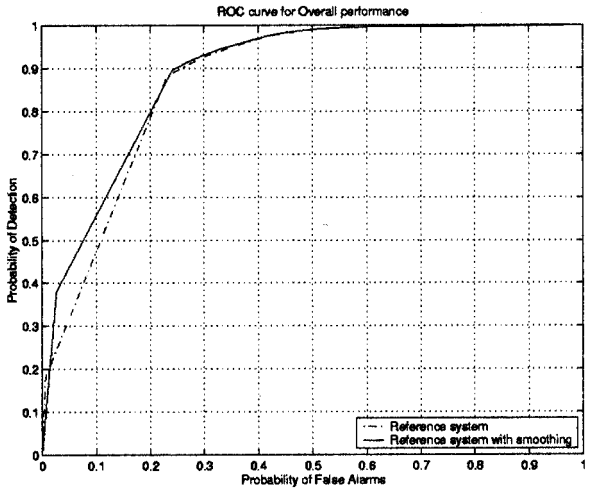


Fig. 13. Comparing the baseline performance of the reference system with and without temporal smoothing.

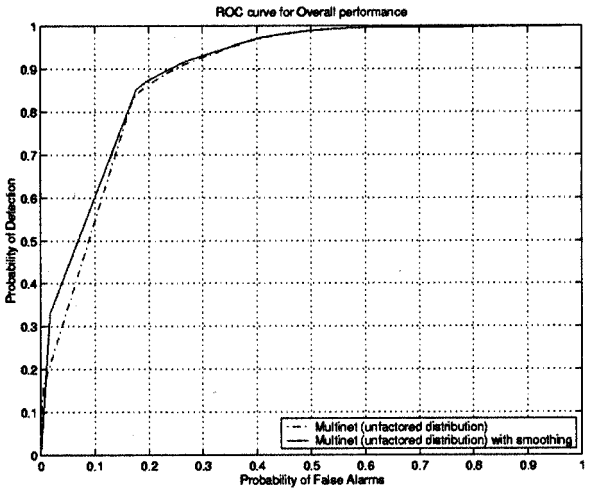


Fig. 14. Comparing the performance of the multinet with unfactored global distribution (Fig. 5) with the dynamic multinet using unfactored global distribution and temporal smoothing [Fig. 7(a)].

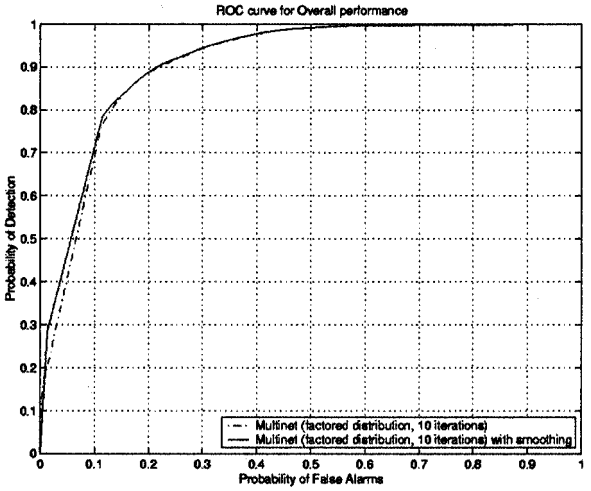


Fig. 15. Comparing the performance of the multinet with factored global distribution (Fig. 6) with the dynamic multinet using factored global distribution and temporal smoothing [Fig. 7(b)].

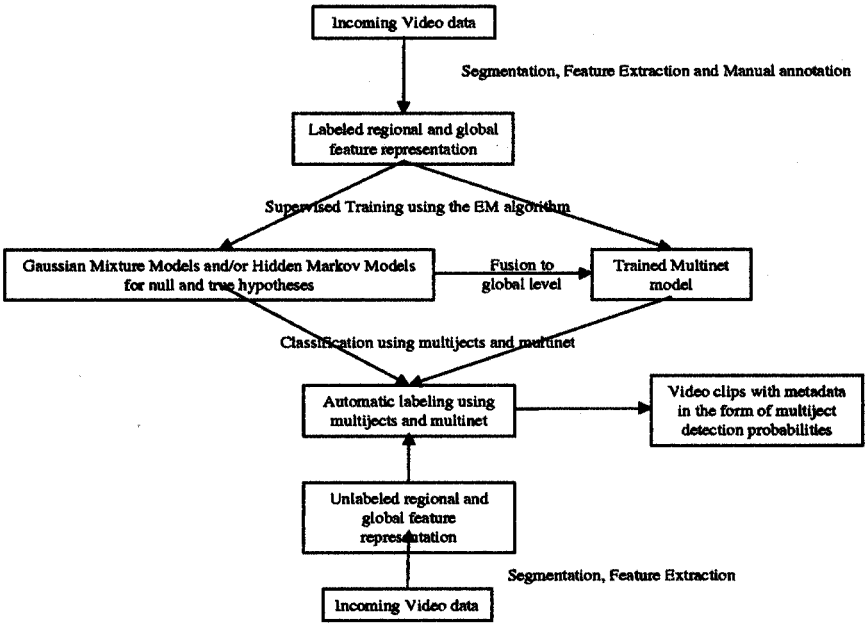


Fig. 16. Block diagram of the entire system.

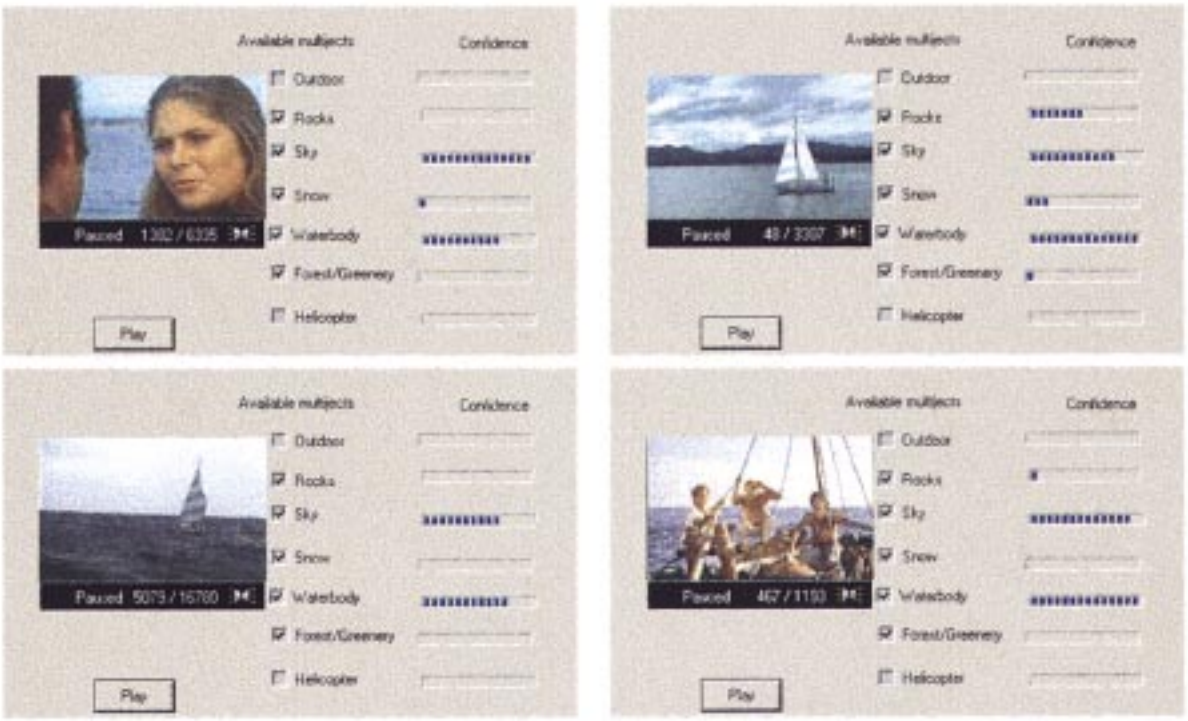


Fig. 17. Four clips retrieved when searched using the keywords *sky* and *water*.

temporal dependencies across video frames and by performing smoothing using the forward backward algorithm. The improvement in performance is upto 9% in Fig. 13, 12% in Fig. 14, and 9% in Fig. 15. Maximum improvement occurs at very low false alarm rates i.e., in the range of $0 \leq P_f \leq 0.05$.

The spatio-temporal context modeled by the multinet, exploits the mutual information among the multijects to enhance detection performance. If the concepts are independent, there will be no gains by modeling the context. On the other hand, as is often observed, if there is mutual interaction between concepts, the multinet will then enforce this interaction and improve detection. This observation holds true irrespective of the nature of the video. The gain in detection is directly related to the amount of inter-dependence between concepts and greater gains in detection are predicted with greater inter-dependence.

C. Filtering and Semantic Indexing Using the Framework

The block diagram for the system using the multijects and multinet for semantic video indexing is shown in Fig. 16. 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. Since the soft

decisions are available the user can vary the threshold for each multijject to tune the filter. 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 keyword search and bring video clips at par with text-databases. Popular internet search engines can definitely be enhanced if they support keyword based video search. Fig. 17 shows four clips retrieved when searched using the keywords *sky* and *water*. Keywords such as *sky, greenery,* and *explosion* that are used for querying represent high-level concepts and the system fetches clips, which contain these concepts with the required degree of confidence, that a user desires. Since we provide the confidence measures, thresholds can be personalized for the individual user. 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.

V. Future Research and Conclusions

In this paper, we have presented a novel probabilistic framework for semantic video indexing. The framework is based on multijects and multinets. We have presented a framework to obtain multiject models for various objects sites and events in audio and video. The procedure remains identical for a large variety of multijects. To discover the relationship and interaction between multiject models, we have presented a factor graph multinet and described how it is automatically learnt. Using the multinet to explicitly model the interaction between multijects, we have demonstrated a substantial improvement in detection performance and also facilitated detection of concepts, which may not be directly observed in the media features. We have also extended the multinet to account for temporal dependencies within concepts across consecutive video frames within shots. This has lead to further performance improvement. 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 support inference of those concepts which may not be observable in the multimedia features. The multinet architecture does not impose any conditions on the multiject architecture except that it be probabilistic. We can, therefore, experiment with more sophisticated class-conditional density functions for modeling multijects. This will lead to an improvement in the baseline performance, as well as system performance. In the future, we will also attempt to model characteristics of the interaction between semantic concepts other than co-occurrence. Spatial layout is one aspect that needs to be modeled together with co-occurrence for better modeling of context.

Acknowledgment

The authors thank D. Zhong and Dr. S. F. Chang for the spatio-temporal segmentation and tracking algorithm [2].

References

[1] J. R. Smith and S. F. Chang, "Visualseek: A fully automated content-based image query system," in *Proc. ACM Multimedia*, Boston, MA, Nov. 1996.

[2] D. Zhong and S. F. Chang, "Spatio-temporal video search using the object-based video representation," *Proc. IEEE Int. Conf. Image Processing*, vol. 1, pp. 21–24, Oct. 1997.

[3] Y. Deng and B. S. Manjunath, "Content based search of video using color, texture and motion," *Proc. IEEE Int. Conf. Image Processing*, vol. 2, pp. 534–537, Oct. 1997.

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

[5] M. M. Yeung and B. Liu, "Efficient matching and clustering of video shots," *Proc. IEEE Int. Conf. Image Processing*, vol. 1, pp. 338–341, Oct. 1995.

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

[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. SPIE Symp.*, vol. 2419, San Jose, CA, Feb. 1995, pp. 1–11.

[9] H. J. Zhang, C. Y. Low, and S. Smoliar, "Video parsing using compressed data," in *Proc. SPIE Conf. Image and Video Processing II*, San Jose, CA, 1994, pp. 142–149.

[10] M. Naphade, R. Mehrotra, A. M. Ferman, J. Warnick, T. S. Huang, and A. M. Takalp, "A high performance shot boundary detection algorithm using multiple cues," *Proc. IEEE Int. Conf. Image Processing*, vol. 2, pp. 884–887, Oct. 1998.

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

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

[13] 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.

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

[15] W. Wolf, "Hidden Markov model parsing of video programs," in *Proc. Int. Conf. Accoustics Signal and Speech Processing*, 1997.

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

[17] N. Vasconcelos and A. Lippman, "Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing," *Proc. IEEE Int. Conf. Image Processing*, vol. 2, pp. 550–555, Oct. 1998.

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

[19] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley Eastern, 1973.

[20] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Its Extensions*. New York, NY: Wiley, 1998.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

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

[23] M. R. Naphade and T. S. Huang, "Semantic filtering of video content," in *Proc. SPIE Storage and Retrieval for Multimedia Databases*, vol. 4315, Jan. 2001, pp. 270–279.
[24] M. R. Naphade, X. Zhou, and T. S. Huang, "Image classification using a set of labeled and unlabeled images," in *Proc. SPIE Photonics East, Internet Multimedia Management Systems*, Nov. 2000.
[25] M. Naphade, I. Kozintsev, T. Huang, and K. Ramchandran, "A factor graph framework for semantic indexing and retrieval in video," in *Proc. Workshop on Content Based Access to Image and Video Labraries Held in Conjunction with CVPR*, June 2000, pp. 35–39.
[26] B. J. Frey, *Graphical Models for Machine Learning and Digital Communications*. Cambridge, MA: MIT Press, 1998.
[27] F. Kschischang, B. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, pp. 498–519, Feb. 2001.
[28] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.
[29] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann, 1988.
[30] R. G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA: MIT Press, 1963.
[31] N. Wiberg, "Codes and Decoding on General Graphs," Ph.D. dissertation, University of Linköping, Linköping, Sweden, 1996.
[32] B. Frey and F. Kschischang, "Probability propagation and iterative decoding," in *Proc. Annu. Allerton Conf. Communication, Control and Computing*, Sept. 1996.
[33] ——, "Probability propagation and iterative decoding," in *Proc. 34th Annu. Allerton Conf. Communication, Control and Computing*. Urbana, IL, 1997.
[34] F. V. Jensson, *Introduction to Bayseian Networks*. New York: Springer Verlag, 1996.
[35] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*. New York: MIT Press and McGraw-Hill, 1995.
[36] A. K. Jain and A. Vailaya, "Shape-based retrieval: A case study with trademark image databases," *Patt. Recognit.*, vol. 31, no. 9, pp. 1369–1390, 1998.
[37] A. K. Jain, A. Vailaya, and W. Xiong, "Query by video clip," *Multimedia Syst.: Special Issue on Video Libraries*, vol. 7, no. 5, pp. 369–384, 1999.
[38] S. Dudani, K. Breeding, and R. McGhee, "Aircraft identification by moment invariants," *IEEE Trans. Comput.*, vol. 26, pp. 39–45, Jan. 1997.

Milind Ramesh Naphade received the B.E. degree in instrumentation and control engineering from the University of Pune, Pune, India, in July 1995, ranking first among the university students in his discipline, and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign (UIUC) in 1998 and 2001, respectively.

He was a Computational Sciences and Engineering Fellow and a member of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology, UIUC, from August 1996 to March 2001. In April 2001, he joined the Pervasive Media Management Group, IBM T. J. Watson Research Center, Hawthorne, NY, as a research Staff Member. He has worked with the Applications Development Group in the Center for Development of Advanced Computing (C-DAC), Pune, India, from July 1994 to July 1996, the Kodak Research Laboratories, Eastman Kodak Company, in the summer of 1997, and also with the Microcomputer Research Laboratories, Intel Corporation, in the summer of 1998. His research interests include audio-visual signal processing and analysis for the purpose of multimedia understanding, content-based indexing, retrieval, and mining. He is interested in applying advanced probabilistic pattern recognition and machine learning techniques to model semantics in multimedia data.

Igor V. Kozintsev received the diploma (Hons.) from the Moscow State Technical University, Bauman, Moscow, Russia, in 1994, and the M.S. and Ph.D. degrees in 1997 and 2000 from the University of Illinois at Urbana-Champaign (UIUC), all in electrical engineering.

During 1996–2000, he was a Research Assistant at the Image Formation and Processing Laboratory, Beckman Institute for Advanced Science and Technology, UIUC. In May 2000, he joined the Technology and Research Labs at Intel Corporation, Santa Clara, CA, where he is currently a Senior Researcher. His research interests include multimedia processing, wireless communications, and networking.

Thomas S. Huang received the B.S. degree 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, all in electrical engineering.

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 at Purdue University, West Lafayette, IN, and Director of its Laboratory for Information and Signal Processing from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign (UIUC), where he is now a William L. Everitt Distinguished Professor of Electrical and Computer Engineering and Research Professor at the Coordinated Science Laboratory. He is also Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology, UIUC. During his sabbatical leaves, he has worked at the MIT Lincoln Laboratory, the IBM Thomas J. Watson Research Center, Hawthorne, NY, and the Rheinishes Landes Museum, Bonn, West Germany. He has held Visiting Professor positions at the Swiss Institutes of Technology of Zurich and Lausanne, the University of Hannover, West Germany, INRS-Telecommunications of the University of Quebec, Montreal, Canada, and the University of Tokyo, 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.

Dr. Huang is a Founding Editor of the International *Journal Computer Vision, Graphics, and Image Processing* and Editor of the Springer Series in Information Sciences. He was the recipient of 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, the Society Award in 1991, and the IEEE Third Millennium Medal and the Honda Lifetime Achievement Award for "contributions to motion analysis" in 2000. He is a Fellow of the International Association of Pattern Recognition of the IEEE and of the Optical Society of America.