RC22580 (W0210-028) October 3, 2002
Other

IBM Research Report

Statistical Media Analysis in Video Indexing

Milind R. Naphade
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Hawthorne, NY 10532

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Statistical Media Analysis in Video Indexing

Milind R. Naphade

IBM Thomas J. Watson Research Center
Pervasive Media Management Group
19 Skyline Drive, Hawthorne, NY 10532
Email: naphade@us.ibm.com

ABSTRACT

Media analysis for video indexing is witnessing an increasing influence of statistical techniques. Examples of these techniques include the use of generative models as well as discriminant techniques for video structuring, classification, summarization, indexing and retrieval. Advances in multimedia analysis are related directly to advances in signal processing, computer vision, pattern recognition, multimedia databases and smart sensors. This paper highlights the statistical techniques in multimedia retrieval with particular emphasis on semantic characterization.

Keywords: Multimedia Analysis, Statistical machine learning, Decision theory, Semantic video indexing, probabilistic graphical models. expectation maximization algorithm, factor graphs

1. INTRODUCTION

The task of automatic analysis is to reduce the tremendous volume of multimodal data to concise representations, that capture the essence of the data. Tools for efficient storage, retrieval, transmission, editing, and analysis of multimedia content are absolutely essential for the utilization of raw content. Such content confronts us everywhere. Several aspects of our daily interaction with the world are affected by the modes of broadcasting, communication, and computing. Video* databases serve as a perfect example of how the acute need for tools has severely constrained the use of multimedia content. One of the greatest challenges is the mapping between computational multimedia features and the high-level semantics represented by the media. Multimedia databases can be better accessed if the index generated contains semantic concepts. Efficient analysis can enable applications such as high-level filtering, automatic rating of Internet sites, restricting access to violent content better and natural interfaces in human computer interaction very low bit-rate video coding, video skimming, summarization, and transcoding etc.

This paper analyzes the application of statistical techniques in the advancement of video indexing. The focus is on statistical techniques for video classification, structuring, summarization, filtering and browsing.

2. REVIEW

In this section the state of the art in image sequence, audio and video retrieval are briefly reviewed.

2.1. Image sequence retrieval

Most techniques for image sequence retrieval support the query by example paradigm.[1,2] These systems represent video clips using characteristics such as color texture, motion, object trajectories etc. and match them using similarity measures like Euclidean distance. Most systems use visual shot boundary detection followed by the grouping of shots to create scenes for easy browsing. Keyframes are extracted to represent shots.

Statistical models like the hidden Markov models (HMM) have been used for structuring image sequences.[3–5] Topical classification of image sequences can provide information about the genres of videos like news, sports, etc. Examples include.[6–8] Extraction of semantics from image-sequences is difficult. Recent work dealing with

video refers to an image sequence and its accompanying audio track.

semantic analysis of image sequences include Naphade et al,[9] and Brand et al.[10] Naphade et al[9] use hidden Markov models to detect events in image sequences. Brand et al.[10] use coupled HMMs to model complex actions in *Tai Chi* movies.

2.2. Audio retrieval

Compared to the significant advances in speech recognition and speech processing related technologies, there has been little progress in terms of retrieval of nonspeech audio data. Most indexing and retrieval schemes in audio deal with human speech (sitcoms, radio interviews, news) with relatively noise-free environment and work on a vocabulary of words. Recent examples include.[11] Recently, there have been attempts to segment the sound track in motion pictures[12] and television comedies.[13] The sound tracks in motion pictures offer tremendous untapped potential in terms of capturing the semantics if intelligent systems can make use of their rich information content. Auditory scene analysis attempts to capture information in the audio track. Two of the most frequently used classes in auditory scene analysis include *speech* and *music*. Recent work in segmentation and classification of audio streams includes.[12–18] Naphade and Huang[12] used hidden Markov models to classify audio streams.

2.3. Video retrieval

While numerous techniques are available for the characterization and retrieval of the individual media mentioned above, techniques using both the modalities for audiovisual analysis are few. Most techniques, using audiovisual data perform temporal segmentation on one medium and then analyze the other medium as described earlier. For example the image sequence is used for temporal segmentation and the audio is then analyzed for classification. Examples include,[11,19] and the Informedia project[20] that uses the visual stream for segmentation and the audio stream for content classification. Such systems also exist for particular video domains like broadcast news,[21] sports,[6,17,22] meeting videos[23] etc. Wang et al.[19] survey a few techniques for analysis using a similar approach for similar domains. In case of domain-independent retrieval, while existing techniques attempt to determine what is going on in the speech-audio, most techniques go as far as classifying the genre of the video using audiovisual features.[8] Other techniques for video analysis include the unsupervised clustering of videos.[24] Naphade et al.[25] have presented an algorithm to support query by audiovisual content. Another popular domain is the detection and verification of a speaker using speech and an image sequence obtained by a camera looking at the person.[26] This is particularly applicable to the domain of intelligent collaboration and human-computer interaction. Recent work in semantic video indexing includes Naphade et al.[9,27,28]

2.4. Level of Analysis: Low-level vs. High-level

Most existing systems fall in the category of low level analysis. This includes most systems supporting the paradigm of query by image/video content through examples or sketches as they use similarity in low-level feature spaces for retrieval. A major drawback with the QBE paradigm is the need for good examples. Unless good examples can initiate the search, good results are difficult to obtain.

An efficient video retrieval system *must* be able to handle the semantics of the query. It must abstract the relationship between low-level media features and high-level semantic concepts to allow the user to query in terms of these concepts rather than in terms of examples. Research in multimedia analysis, has witnessed a recent shift of interest from syntactic analysis to semantic analysis. This is reflected by the attention given to the semantic description scheme in the emerging MPEG-7 standard.[29]

Systems for video analysis beyond low-level features aim at one of the following objectives.

- Detecting high-level structures like dialogs and scenes, commercials[8] and other such recurring patterns common to multiple domains.[3,4,12]
- Classifying the genre or the topic of the video clip. Examples include the MoCA system,[8] and Kobla et al.[6] Such categorization has its roots in similar approach for topical detection from text.
- Domain-dependent analysis like that of sports video (football[30]), and news video.[20] In such systems, the knowledge is built manually by the human expert in the form of rules.

- Domain-independent analysis for automatic annotation of video. This remains the **final frontier** in multimedia analysis for retrieval. This necessitates the detection of semantic concepts in video.[9,31–34] Naphade et al.[9,31] present novel ideas in semantic video indexing by learning probabilistic multimedia representations of semantic concepts, including semantic events like *explosion* and *waterfall*. Semantic classification schemes include those for images[35–37] as well as those for image sequences.[8,9,34]

2.5. Multiple Modalities and Fusion

Using information available from multiple sensors is a challenging fusion problem. Video analysis presents a challenge in terms of the fusion of audio visual and textual modalities. No single modality is sufficient to capture representations especially for semantic concepts. For example, it is difficult to recognize (detect) a helicopter from an image sequence alone. However, detecting the sound of the helicopter may not be as difficult in audio. This example illustrates the importance of going beyond vision or speech to address multimedia content in a unified fashion. It is not clear how information available from multiple modalities should be fused. Approach to efficient fusion often depends on the nature of audiovisual interaction as well as the desired end-result of the analysis. The information in these modalities may corroborate or contradict. The degree and type of correlation between the multiple media streams varies with the concept being conveyed. Of the several techniques referred to in this paper, only a few can claim to be truly multimodal in terms of joint utilization of audio-visual data. Examples include.[8,21,24–27,31] Most systems supporting query by example support unimodality. The multimodal QBE system presented by Naphade et al[25] for example, is one of the very few QBE systems that are truly multimodal.

2.6. The Role of the User in the Loop

Since the aim of any retrieval, indexing, filtering, and browsing system is to aid the human, the involvement of the human in the process is crucial. Human-centric applications such as multimedia content management can benefit largely if the system can figure out when human intervention is crucial and how it should be optimally utilized.

Existing systems can thus be differentiated based on how the human is involved in the system. A possible classification is based on the stage at which the human is involved. Systems supporting relevance feedback and browsing involve the human at a later stage during the retrieval/browsing process. On the other hand systems can involve the human at an earlier stage for the process of annotation (as in supervised training). Human intervention is considered costly in most cases. Thus the effort of the system is to adapt and minimize the user's interaction. In systems that need the human for annotation, techniques such as active learning are being used[38] to select those examples that are the most informative if annotated. In relevance feedback systems where the user is involved in an iterative procedure the main aim is to minimize the number of iterations. Yet another approach is to use the user annotated examples along with the unlabeled examples to drive the process of classification.[35,39]

3. STATISTICAL METHODS FOR DEFINING SIMILARITY

Probabilistic representation of features implies the use of probability density functions to represent the statistical properties of features. In the QBE paradigm, similarity in media feature spaces is usually defined in terms of Euclidean or Mahalanobis distance. However if these features are represented as probability density functions, then measures such as the Kullback Leibler divergence[40] can be used to define similarity. Examples of probabilistic matching include[37,41,42] etc. Vasconcelos[42] compares various statistical measures and their complexity. Amongst other things, Vasconcelos shows how a query image can be matched with the images in a database with less computational complexity by using probabilistic representations and statistical distance measures. Greenspan et al.[41] use mixtures of Gaussians to represent dominant regions in images and extend the model for image sequences.

4. STATISTICAL METHODS FOR SEMANTIC CLASSIFICATION

For automatic semantic video indexing and retrieval, models representing semantics in terms of multimedia features are needed. This can be viewed as a multimedia pattern recognition problem. Depending on the nature of the semantic concept, the model and the training algorithm can be chosen. Most of the semantic concepts to be modeled can be classified either as objects, sites or events. Sites represent static concepts like *sky*, *mountain*, *outdoor*, *cityscape*, etc. Examples of events include *explosion*, *gunshot*, *dancing*, *helicopter-flying*, etc. Events are those semantic concepts, which evolve temporally. Examples of objects include *face*, *car*, *animal*, *building* etc.

This classification of semantic concepts is not rigorous, but helps in choosing models for the three concept classes. For example, it is clear that sites should be modeled with static pattern recognition techniques with long-term temporal support only for the sake of visual continuity. Similarly it is evident that events need the short-term temporal dynamics to be modeled and this is a time-series classification problem. Objects on the other hand point to definitive shape or structure and it is easier to model them if they are associated with events. For example, it is easier to model a flying helicopter or a moving car than it is to model a static car or helicopter from shape. The principal difficulty is imperfect segmentation of object shape from the background. Another reason for favoring detection of events related to objects over detecting the objects statically, is the hope to see some characterization in the audio track. This serves as an additional source of information, often loosely coupled with the visual characterization.

4.1. Supervised Learning and Probabilistic Feature Representation

A component for model-based learning/recognition is necessary in order to develop representations for semantic concepts. Supervised or partially supervised training then becomes an inherent component for developing concise representations of semantic concepts that can be used efficiently for semantic retrieval. A desirable characteristic of such a learning/recognition component is the ability to account for the uncertainty in the information in multimedia representation.

Assume that features from audiovisual streams have been computed. Let X be the feature vector. Assume that the statistical properties of these features are characteristic. For distinct instances of all multijects, further assume, that these features are independent identically distributed random variables drawn from known probability distributions, with unknown deterministic parameters. For the purpose of classification, assume that the unknown parameters are distinct under different hypotheses and can be estimated. In particular, each semantic concept is represented by a binary random variable. The two hypotheses associated with each such variable are denoted by H_i, $i \in \{0,1\}$, where 0 denotes absence and 1 denotes presence of the concept. Under each hypothesis, assume that the features are generated by the conditional probability density function $P_i(X)$, $i \in \{0,1\}$. In case of site multijects, the feature patterns are static and represent a single frame. In case of events, with spatiotemporal support, X represents a time series of features over segments of the audiovisual data. The *one-zero* loss function is used[43] to penalize incorrect detection. This is shown in Equation 1:

$$\begin{aligned} \lambda(\alpha_i|\omega_j) &= 0 \quad i = j \\ &= 1 \quad i \neq j \end{aligned} \tag{1}$$

The risk corresponding to this loss function equals the average probability of error and the conditional risk with action α_i is $1 - P(\omega_i|x)$. To minimize the average probability of error, that class ω_i must be chosen, which corresponds to the maximum a posteriori probability $P(\omega_i|x)$. This is the minimum probability of error (MPE) rule.

In the special case of binary classification, this can be expressed as deciding in favor of ω_1 if

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{(\lambda_{21} - \lambda_{11})P(\omega_1)} \tag{2}$$

The term $p(x|\omega_j)$ is the *likelihood* of ω_j and the test based on the ratio in Equation (2) is called the *likelihood ratio test*.[43, 44]

Recent work in modeling concepts include the framework of multijects by Naphade et al.,[9,27,32,35] classification of vacation images by Vailaya et al.,[37] semantic visual templates by Chang et al.,[2] classification of video captured by a wearable camera by Clarkson and Pentland,[24] and classification of soundtrack into meaningful classes such as speech, music.[12,14,17]

In the multijects framework (see Naphade et al[9]) a video shot can be explained as an object or an event occurring at a site or location.

A multiject represents a semantic concept that is supported by multiple media features at various levels (low level, intermediate level, high level) through a structure that is probabilistic.[9] Multijects belong to one of the three categories: objects (*car*, *man*, *helicopter*), sites (*outdoor*, *beach*), or events (*explosion*, *man-walking*, *ball-game*). Figure 1 illustrates the concepts of a multiject.

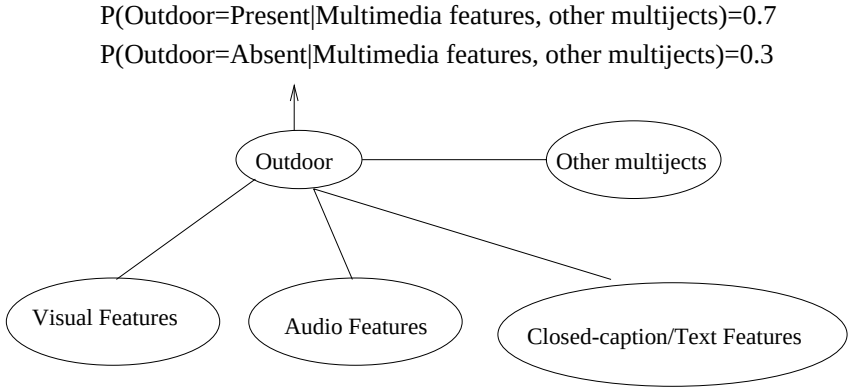

Figure 1. A probabilistic multimedia object (multiject).

A multiject is a flexible, open-ended semantic representation. It draws its support from low-level features of multiple media including audio, image, text, and closed caption.[9] It can also be supported by intermediate-level features, including semantic templates.[2] It can also use specially developed high-level feature detectors like face detectors or other multijects. A multiject can be developed for a semantic concept if there is some correlation between low-level multimedia features and high-level semantics. In the absence of such correlation, a sufficiently invariant representation cannot be learnt. Fortunately many semantic concepts are correlated to some multimedia features, and so the framework has the potential to scale.

4.1.1. Site multijects based on image sequences

In case of sites the feature vector is modeled as a Gaussian mixture model (GMM). The temporal flow is not taken into consideration. The expectation maximization (EM) algorithm[45] is used to estimate the means and covariance matrices of the GMMs. Sites may be supported by blobs or regions, or they may be supported by the whole frame. Several regional site multijects have been modeled *rocky-terrain, sky, snow, water-body, forest/greenery, outdoor*[46] etc. .

4.1.2. Event multijects based on audio

Interesting semantic events in audio include *speech*, *music*, *explosion*, *gunshots*, etc. Recent work in statistical audio modeling includes.[12–16,18] Naphade and Huang[12] used hidden Markov models (HMMs) for representing the probability density functions of auditory features computed over a time series. An audio event is modeled using a set of states with a Markovian state transition and a Gaussian mixture observation density in each state. Continuous density models can be used in which each observation probability distribution is represented by a mixture density. For state j the probability $b_j(\mathbf{o_t})$ of generating observation $\mathbf{o_t}$ is given by Equation (3):

$$b_j(\mathbf{o_t}) = \sum_{\mathbf{m=1}}^{\mathbf{M_j}} \mathbf{c_{jm}} \mathcal{N}(\mathbf{o_t}; \mu_{\mathbf{jm}}, \mathbf{\Sigma_{jm}}) \tag{3}$$

where M_j is the number of mixture components in state j, c_{jm} is the weight of the m^{th} component, and $\mathcal{N}(\mathbf{o}; \mu, \mathbf{\Sigma})$ is the multivariate Gaussian with mean μ and covariance matrix $\mathbf{\Sigma}$. The parameters to be estimated are the transition matrix A, the mixing proportions c, and the observation densities b. The Baum-Welch reestimation procedure[47,48] is used to train the model and estimate the set of parameters. Once the parameters are estimated using the training data, the trained models can then be used for classification as well as state sequence decoding.[47,48] Some of the multijects developed include *human-speech*, *music*[12,49] and *helicopter-flying*.[27]

The audio stream in motion picture soundtrack is of a composite nature. This corresponds to the mixing of sounds from different sources. Speech in foreground and music in background are common examples. The coexistence of multiple individual audio sources forces us to model such events explicitly. If sufficient data for training is available, HMM models (based on audio alone) that represent concurrent occurrence of multiple multijects[12] can be estimated.

4.1.3. Event multijects based on video

Most probabilistic techniques for modeling features from modalities having temporal support are based on Markov models. Examples include the HMM[48] and its several variants for fusing multiple modalities like the coupled HMM,[10] factorial HMM,[50] etc. Fusion of multimodal feature streams (especially audio and visual feature streams) has been applied to problems like bimodal speech,[51] summarization of video,[21] query by audiovisual content,[25] and event detection in movies.[9] These models are characterized by the stage at which the features from the different modalities are merged. Examples of audiovisual events include *explosions*, *human-talking*, etc.

Assuming synchronization, the two main categories of fusion models are those that favor early integration of features as against those that favor late integration. Models for early integration include the coupled HMM,[10] the factorial HMM[50] etc. The audio and visual streams in a movie or a news clip appear to be loosely coupled. Late integration includes schemes, which use independent models for multiple feature streams and then combine the weighted decisions. Naphade et al[9] have developed a hierarchical HMM (HHMM) that combines features with temporal support from multiple modalities and performs late integration. Under each hypothesis it is assumed that each feature stream is generated by an HMM.[48] The parameters of the HMM under each hypothesis for each feature stream are estimated using the EM algorithm[45,48] and a labeled training set. The best state sequence and hypothesis for the test set is then evaluated through maximum-likelihood detection. Once the state sequence for each feature stream is obtained, these intermediate-level decisions are used as features for a supervisor HMM,[9] that combines these intermediate level decisions for all feature streams.

Figures 2 shows an HHMM with audio and image sequence features forming the two media observations streams.

Each state in the media HMMs represents a stationary distribution and by using the Viterbi decoder over each feature stream, features are clustered spatiotemporally and quantized through state identities.[9] The optimal state sequences obtained using the Viterbi algorithm[48] in the video and audio are now treated as the observations of the supervisor HMM. The supervisor HMM encodes the correlation of states in the two modalities. This is a hierarchical, fast, greedy bottom-up algorithm.

4.1.4. Complex action recognition using coupled HMMs

Brand et al.[10] present a coupled HMM for detecting complex actions in Tai Chi movies. The multiple sensors in this situation are the location and movement of the two arms of the performer. Brand et al.[10] argue that these movements are not independent for Tai Chi actions and model the interaction through inter-dependence of state sequences of two hidden Markov models. This coupling leads to an improvement in performance over linked HMMs and individual HMMs. In a coupled HMM, there are as many hidden state chains as number of feature streams and at each time instance in each chain, transition to a state depends on the identity of the previous state of the same chain as well as the previous state in other chains. The $3D$ hand tracking data is obtained by using a self-calibrated stereo blob tracker.

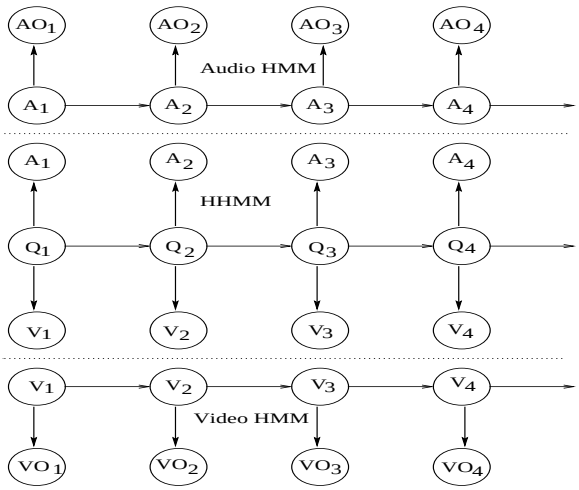

Figure 2. Hierarchical HMM: The state sequence within the dotted lines represents the supervisor HMM sequence. Nodes $VO_1, \ldots, VO_m$ denote video observations. Nodes $AO_1, \ldots, AO_m$ denote audio observations. Nodes $V_1, \ldots, V_m$ denote the decoded state sequence for the video HMM. Nodes $A_1, \ldots, A_m$ denote the decoded state sequence for the audio HMM. Nodes $Q_1, \ldots, Q_m$ denote the state sequence for the supervisor HMM.

4.1.5. Genre Detection

A popular application of statistical techniques has been for detection of genres in videos. Examples include[6–8, 52] etc. Popular classes include *sports, news, commercials, television shows* etc. Detailed classification of sports into baseball, football etc. has also been studied.

5. STATISTICAL METHODS FOR REPRESENTATION OF KNOWLEDGE AND CONTEXT

Humans analyze multimedia content and its associated semantics using the audiovisual sensory information as well as knowledge learnt over an extended length. The knowledge base is used for reasoning and inference. It is obvious that semantics is meaningful only in context and not in vacuum. This fact has been used in information retrieval in the text domain by WordNet.[53] WordNet is a text-retrieval system built manually over several years that encodes relationships of different kinds between the words supported in the system. In speech recognition and natural language processing stochastic grammars have been proposed to improve the detection performance.

To incorporate domain knowledge we need to influence the decisions by context, grammar and other high-level information. This information may be domain dependent. The constraints may be represented by rules heuristics or association. The emerging MPEG-7 standard on multimedia content description, attempts to define a language for semantic description[29] to describe interactions that are possible between a pair of semantic concepts.

Ideally, we would like to learn the context from data rather than having to build it manually into the system. Neural nets, expert systems, fuzzy logic, decision trees, static and dynamic Bayesian networks, factor graphs, Markov random fields, etc. are popular mechanisms for storing and enforcing high-level information. For example the high-level analysis may use classifiers of *sky*, *rocks*, *snow*, *water-body*, *forestry*, etc. to infer that the scene is an *outdoor* scene.[54] The enforcement of constraints due to grammar has been shown to improve detection performance in automatic speech recognition.

Recent work in computational models for contextual relations between semantic concepts included a probabilistic graphical framework (*multinet*).[9, 55] The multinet framework[9, 55] is one of the earliest framework for learning computational models of contextual association and constraints between several semantic concepts from

an annotated training data. A multinet is a network of distributed semantic concepts, linked through edges which denote the nature and the strength of the spatio-temporal interaction between these semantic concepts.

Semantic concepts do not occur independently or in isolation from each other. In fact significant information lies in their co-occurrence. Intuitively, it is clear that the presence of certain semantic concepts suggests a high possibility of presence of certain others. Similarly, some concepts are less likely to occur in the presence of others. For example, a flying helicopter is mostly shown against sky as the background. Thus sky and helicopter share a positive relation. Similarly, *beach*, *sand*, *sky* and *water* share a positive relation. *Explosion* and *gunshots* similarly share a positive relation with *violence*. Concepts also tend to share negative relations. For example *sky* and *snow* rarely co-occur with *indoor*. There also are spatiotemporal relations. For example *sky* must occur above *water*. When a person speaks, the mouth movements must be synchronous with speech.

The multinet reflects the dependencies, relationships, and co-occurrence between semantic concepts at various hierarchical levels. A multinet provides a robust framework for supporting four aspects of constructing semantic indices:

- Enhancing detection: The use of mutual information can enhance detection of multijects.
- Support inference: Some multijects may not provide the required degree of invariance in feature spaces. To detect such multijects, the multinet can support inference based on the interaction of these multijects with other multijects (which can be detected with greater ease).
- Imposing 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 *gunshots* and *explosion*.
- Combining classifiers: The multinet can combine classifiers and fuse multiple modalities.

A graphical representation of a multinet is shown in Figure 3 with the positive signs indicating a positive interaction and negative signs indicating a negative interaction.

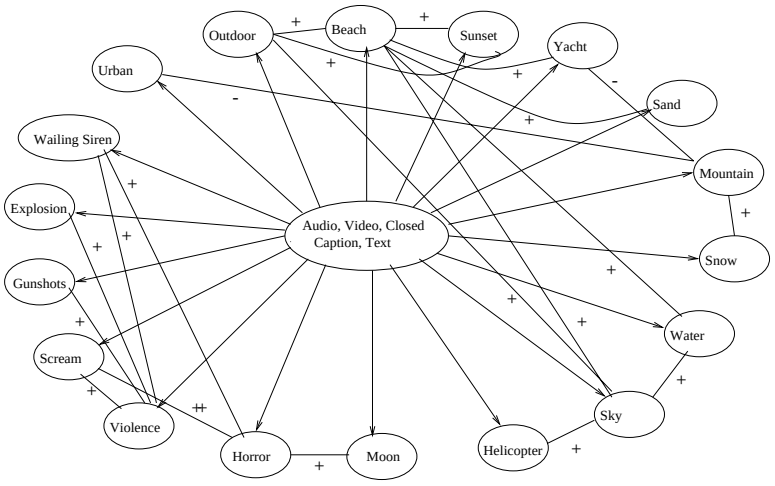

Figure 3. A conceptual figure of a multinet. Signs indicate positive or negative interaction between concepts.

Contextual constraints can change with time and with domain. Dynamic multinets, that can take care of such changes are therefore necessary. Naphade et al[46,55] presented a probabilistic graphical network architectures as a computational model of the multinet. The backbone of this computational model is the *factor graph* framework.[55–57] Using the sum-product algorithm, we are able to propagate evidence in factor graphs and perform learning as well as inference. For details see Naphade et al.[55]

Applying the theory of multijects and multinet, a system for semantic filtering was developed for a small set of concepts.[55] An example of the semantic filtering system for movie clips is shown in Figure 4.

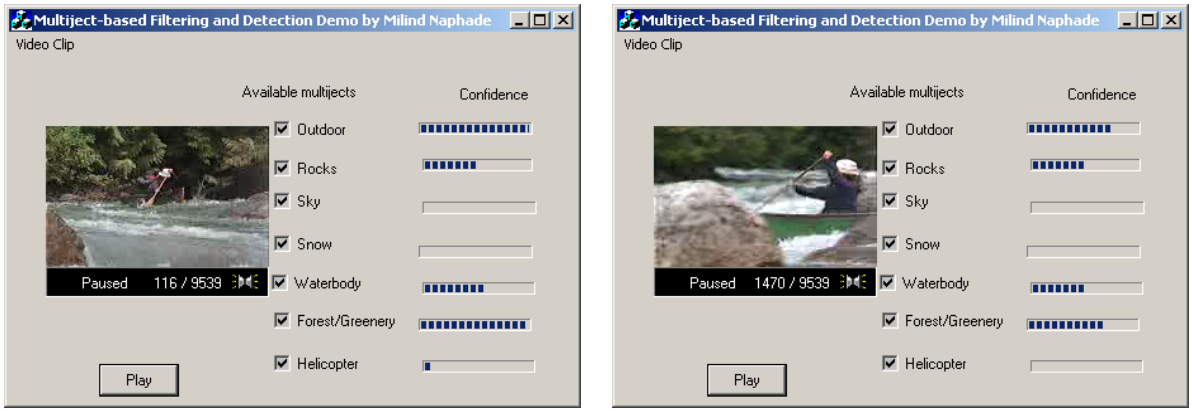

Figure 4. Key frames from a movie and the corresponding confidence measures for various semantic labels.

6. STATISTICAL METHODS FOR SEGMENTATION STRUCTURING AND GROUPING

Probabilistic models like the hidden Markov model and its variants have been popularly used for the temporal segmentation, clustering and grouping of videos. Examples include segmentation and structuring such as detection of dialogs[58] and scenes, and other such recurring patterns common to multiple domains.[3,4,12]

Clarkson and Pentland[24] analyze the audio-visual signals from a wearable camera and a microphone. The system captures video while accompanying a person in daily activities. Depending on the audiovisual features, the attempt is to group the data into clusters, which may represent types of activities. Since the approach is unsupervised, this may not always correspond to distinct semantic activities. The unsupervised clustering is then evaluated using a baseline classification system. The baseline system is trained in a supervised fashion as the wearer of the camera and microphone labels the locations while the data is obtained. The different kinds of events for which the models are trained including entering and leaving an office location, a kitchen and a black couch area.

Liu et al.[52] use an ergodic hidden Markov model to cluster audio and then use this information to drive a supervised training scheme for the classification of videos. Naphade and Huang[5] use hierarchical model with non-ergodic hidden Markov models embedded within a long term model which is ergodic. Application of this framework to the discovery of recurring patterns in movies leads to the automatic clustering and detection of events in video such as *explosion*. Application to programs such as talk shows leads to the automatic detection of events such as *monologue*, *laughter*, *applause*, *music* etc.

7. STATISTICAL METHODS FOR USER INTERACTION

From the application view point, user interaction comes into picture in video indexing and retrieval systems in one of the two scenarios. In the first, a user annotates and adds value to a video. In the second, the user changes the parameters of the system to enhance performance according to the user's preference. In the former scenario, the onus is on the system to minimize the input from the user as this interaction is costly and time consuming. In the latter scenario, the biggest problem is the user's impatience in persisting with feedback over long periods of time.

The aim in either case is to minimize the number of inputs from the user while maximizing the accuracy of automatic propagation based on the learnt representations. Two aspects that need to be learnt are the

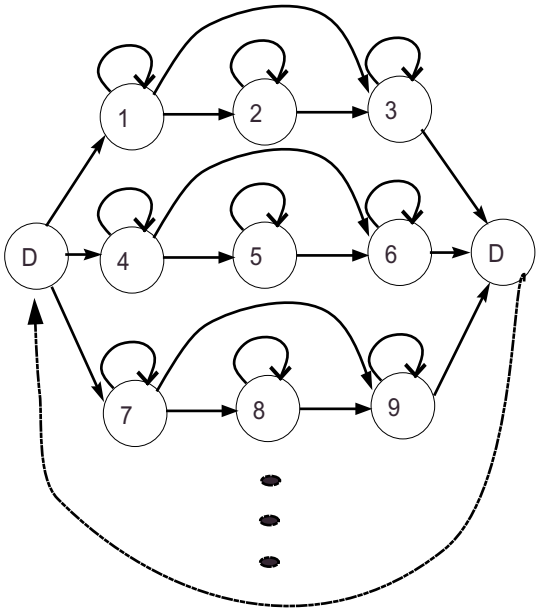

Figure 5. Imposing Structure using several non ergodic hidden Markov models embedded in a hierarchical ergodic configuration.[5] States labeled D are dummy states that do not emit any observations. Each branch is a non ergodic HMM.

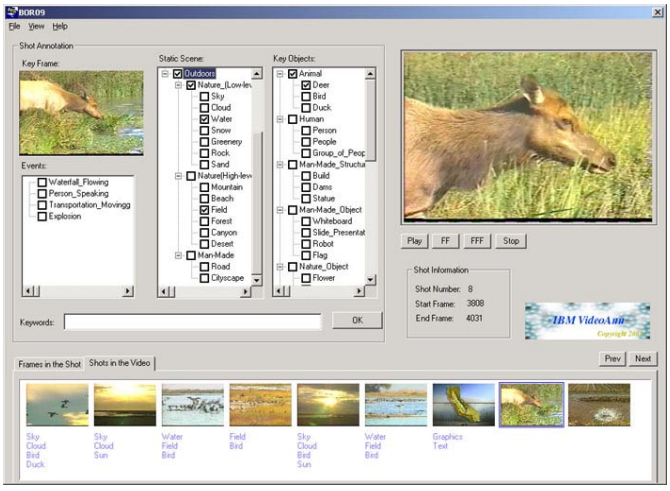

Figure 6. A System for Video Annotation that uses active sample selection to minimize user interaction.

model that can be used to represent the user's inputs and the model that can be used to represent the user's preferences. Again this categorization is purely driven by applications.

Annotation can be assisted through intelligent modeling of the user inputs and through propagation and sample selection procedures. Naphade et al[38] demonstrate that the number of samples that need to be annotated can be cut down by a factor of 10 if the samples to be annotated by the user are actively selected. A support vector machine based active learner is incorporated for the sample selection.

Figure 6 shows a visual interface for the video annotation.[38]

A different approach to reducing the effort of the user that can be coupled with the above active learning

based annotation is the use of unlabeled samples along with the labeled samples to improve the performance of automatic propagation of annotations. Naphade et al.[35] use an EM algorithm and show improvement in performance if sample selection precedes the labeled-unlabaled learning step.

The application of learning in the retrieval step to assist the user and minimize the iterations of feedback has also been studied. Most notably Tong and Chang[59] show substantially improved performance using support vector machine based active learning as a feedback mechanism in image retrieval.

8. CONCLUDING REMARKS AND FUTURE DIRECTIONS

Statistical techniques are increasingly popular in video data management. In particular these techniques are applied in problems like semantic video classification, genre detection, segmentation, summarization, clustering, browsing, structuring and retrieval. In case of semantic analysis, the difficulty lies in bridging the gap between low-level media features and representations that can be computed and the high-level semantic labels. Statistical machine learning techniques play an important role in modeling semantic concepts, knowledge and context, segmentation and grouping, and user interaction.

While the problem of small sample statistics limits the use of traditional techniques, innovations such as labeled and unlabeled learning, active learning and discriminant techniques have made it more feasible to use statistical models for various video indexing problems. The main challenge in future is to attain performance that is considered useful by the end users of the systems.

REFERENCES

1. M. R. Naphade, M. M. Yeung, and B. L. Yeo, "A novel scheme for fast and efficient video sequence matching using compact signatures," in *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, **3972**, pp. 564–572, Jan. 2000.
2. D. Zhong and S. F. Chang, "Spatio-temporal video search using the object-based video representation," in *Proceedings of IEEE International Conference on Image Processing*, **1**, pp. 21–24, (Santa Barbara, CA), Oct. 1997.
3. W. Wolf, "Hidden Markov model parsing of video programs," in *Proceedings of International Conference on Accoustics Signal and Speech Processing*, 1997.
4. A. M. Ferman and A. M. Tekalp, "Probabilistic analysis and extraction of video content," in *Proceedings of IEEE International Conference on Image Processing*, (Kobe, Japan), Oct. 1999.
5. M. Naphade and T. Huang, "Discovering recurring events in video using unsupervised methods," in *IEEE International Conference on Image Processing*, (Rochester, NY), 2002.
6. V. Kobla, D. DeMenthon, and D. Doermann, "Identifying sports video using replay, text and camera motion features," in *Proceedings of SPIE Storage and Retrieval for Media Databases*, **3972**, pp. 332–343, (San Jose, Ca), Jan. 2000.
7. G. Iyengar and A. Lippman, "Models for automatic classification of video sequences," in *SPIE Conference on Storage and Retrieval for Still Image and Video Databases*, pp. 216–227, 1998.
8. S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *Proceedings of ACM Multimedia*, pp. 295–304, (San Francisco, CA), April 1995.
9. 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 *Proceedings of IEEE International Conference on Image Processing*, **3**, pp. 536–540, (Chicago, IL), Oct. 1998.
10. M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceedings of Computer Vision and Pattern Recognition*, pp. 994–999, 1997.
11. J. Nam, A. Cetin, and A. Tewfik, "Speaker identification and video analysis for hierarchical video shot classification," in *Proceedings of IEEE International Conference on Image Processing*, **2**, pp. 550–555, (Santa Barbara, CA), Oct. 1997.
12. M. R. Naphade and T. S. Huang, "Stochastic modeling of soundtrack for efficient segmentation and indexing of video," in *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, **3972**, pp. 168–176, Jan. 2000.

13. M. Akutsu, A. Hamada, and Y. Tonomura, "Video handling with music and speech detection," *IEEE Multimedia* **5**(3), pp. 17–25, 1998.
14. D. Ellis, *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT, Cambridge, MA, 1996.
15. P. Jang and A. Hauptmann, "Learning to recognize speech by watching television," *IEEE Intelligent Systems Magazine* **14**(5), pp. 51–58, 1999.
16. E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification search and retrieval of audio," *IEEE Multimedia* **3**(3), pp. 27–36, 1996.
17. T. Zhang and C. Kuo, "An integrated approach to multimodal media content analysis," in *Proceedings of SPIE, IS&T Storage and Retrieval for Media Databases*, **3972**, pp. 506–517, (San Jose, Ca), Jan. 2000.
18. Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *VLSI Signal processing Systems for Signal, Image and Video Technology* **20**, pp. 61–79, Oct. 1998.
19. Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using audio and visual information," *IEEE Signal Processing Magazine* **17**, pp. 12–36, Nov. 2000.
20. H. Wactlar, T. Kanade, M. Smith, and S. Stevens, "Intelligent access to digital video: The informedia project," *IEEE Computer Digital Library Intiaive special issue* , May. 1996.
21. Y. Nakamura and T. Kanade, "Semantic analysis for video contents extraction - spotting by association in news video," in *Proceedings of ACM International Multimedia Conference*, Nov. 1997.
22. D. D. Saur, Y. P. Tan, S. R. Kulkarni, and P. J. Ramadge, "Automated analysis and annotation of basketball video," in *Proceedings of SPIE Symposium*, **3022**, pp. 176–187, 1997.
23. J. Foote, J. Boreczky, and L. Wilcox, "Finding presentations in recorded meetings using audio and video features," in *IEEE International Conference on Speech Accoustics and Signal Processing*, pp. 3029–3032, 1999.
24. B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *Proceedings of IEEE International Conference on Accoustics Speech and Signal Processing*, (Phoenix, AZ), 1999.
25. M. R. Naphade, R. Wang, and T. S. Huang, "Supporting audiovisual query using dynamic programming," in *ACM Multimedia*, pp. 411–420, Oct. 2001.
26. J. Rehg, K. Murphy, and P. Fieguth, "Vision-based speaker detection using bayesian networks," in *Proceedings of Computer Vision and Pattern Recognition*, **2**, pp. 110–116, (Fort Collins, CO), June 1999.
27. M. R. Naphade and T. S. Huang, "Recognizing high-level audio-visual concepts using context," in *IEEE International Conference on Image Processing*, **3**, pp. 46–49, (Thessaloniki, Greece), Oct 2001.
28. M. R. Naphade, A. Garg, and T. S. Huang, "Duration dependent input output Markov models for audio-visual event detection," in *IEEE International Conference on Multimedia and Expo*, (Tokyo, Japan), August 2001.
29. ISO/IEC JTC 1/SC 29/WG 11/N3966, *Text of 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5 Multimedia Description Schemes*, Final Committee Draft (FCD) ed., March 2001.
30. J. Courtney, "Automatic video indexing via object motion analysis," *Pattern Recongition* **30**, pp. 607–625, April 1997.
31. M. R. Naphade, *A Probabilistic Framework for mapping audio-visual features to high-level semantics in terms of concepts and context*. PhD thesis, University of Illinois at Urbana-Champaign, Department of Electrical and Computer Engineering, May 2001.
32. M. Naphade, I. Kozintsev, T. Huang, and K. Ramchandran, "A factor graph framework for semantic indexing and retrieval in video," in *Proceedings of Workshop on Content Based Access to Image and Video Libraries Held in Conjunction with CVPR*, pp. 35–39, (Hilton Head, SC), June 2000.
33. M. R. Naphade, I. Kozintsev, and T. S. Huang, "On probabilistic semantic video indexing," in *Proceedings of Neural Information Processing Systems*, **13**, pp. 967–973, (Denver, CO), Nov. 2000.
34. R. Qian, N. Hearing, and I. Sezan, "A computational approach to semantic event detection," in *Proceedings of Computer Vision and Pattern Recognition*, **1**, pp. 200–206, (Fort Collins, CO), June 1999.

35. M. R. Naphade, X. Zhou, and T. S. Huang, "Image classification using a set of labeled and unlabeled images," in *Proceedings of SPIE Photonics East, Internet Multimedia Management Systems*, **4210**, pp. 13–24, (Boston, MA), Nov. 2000.
36. A. Ratan, O.Maron, W. Grimson, and T. Lozano-Perez, "A framework for learning query concepts in image classification," in *Proceedings of Computer Vision and Pattern Recognition*, **1**, pp. 423–429, (Fort Collins, CO), June 1999.
37. A. Vailaya, A. Jain, and H. Zhang, "On image classification: City images vs. landscapes," *Pattern Recognition* **31**, pp. 1921–1936, Dec. 1998.
38. M. Naphade, C. Lin, J. Smith, B. Tseng, and S. Basu, "Learning to annotate video databases," in *SPIE Conference on Storage and Retrieval on Media Databases*, (San Jose, CA), 2002.
39. B. Shahshahani and D. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing* **32**, pp. 1087–1095, Sep. 1994.
40. S. Kullback, *Information Theory and Statistics*, Dover Publications, New York, 1968.
41. H. Greenspan, J. Goldberger, and L. Ridel, "A continuous probabilistic framework for image matching," *Computer Vision and Image Understanding* **1**, Dec 2001.
42. N. Vasconcelos, "On the complexity of probabilistic image retrieval," in *International Conference on Computer Vision*, (Vancouver, Canada), 2001.
43. H. V. Poor, *An Introduction to Signal Detection and Estimation,*, Springer-Verlag, New York, 2 ed., 1999.
44. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley Eastern, New York, 1973.
45. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Proceedings of the Royal Statistical Society* **B**(39), pp. 1–38, 1977.
46. M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering and retrieval," *IEEE Transactions on Multimedia, special issue on Multimedia over IP* **3**, pp. 141–151, Mar. 2001.
47. L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics* **37**, pp. 1559–1563, 1966.
48. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings IEEE* **77**, pp. 257–286, Feb. 1989.
49. M. R. Naphade, R. Wang, and T. S. Huang, "Classifying motion picture soundtrack for video indexing," in *IEEE International Conference on Multimedia and Expo*, (Tokyo, Japan), AUgust 2001.
50. Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," *Machine Learning* **29**, pp. 245–273, 1997.
51. T. Chen and R. Rao, "Audio-visual integration in multimodal communication," *IEEE Proceedings* **86**(5), pp. 837–852, 1998.
52. Z. Liu, J. Huang, and Y. Wang, "Classification of tv programs based on audio information using hidden markov model," in *Multimedia Signal Processing*, 1998.
53. G. Miller, "WordNet: A lexical database for english," *Communications of ACM* **38**, pp. 39–41, Nov 1995.
54. M. R. Naphade and T. S. Huang, "Inferring semantic concepts for video retrieval," in *Proceedings of IEEE International Conference on Image Processing*, (Vancouver, Canada), Sep. 2000.
55. M. R. Naphade, I. Kozintsev, and T. S. Huang, "A factor graph framework for semantic video indexing," *IEEE Transactions on Circuits and Systems for Video Technology* **12**, pp. 40–52, Jan 2002.
56. B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.
57. F. Kschischang, B. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory* **47**(2), pp. 498–519, 2001.
58. H. Sundaram and S. F. Chang, "Video scene segmentation using audio and visual features," in *Proceedings of IEEE International Conference on Multimedia and Expo*, **1**, (New York), July 2000.
59. S. Tong and E. Chang, "Support vector machine active learning image retrieval," in *ACM Multimedia*, pp. 107–118, (Canada), 2001.