On supervision and statistical learning for semantic multimedia analysis

Milind R. Naphade

IBM Thomas J. Watson Research Center, Pervasive Media Management Group, 19 Skyline Drive, Hawthorne, NY 10532, USA

Received 30 June 2003; accepted 5 April 2004

Available online 23 August 2004

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. There is increasing emphasis on reducing the amount of supervision and user interaction needed to construct and utilize the semantic models. This paper highlights the statistical learning techniques in semantic multimedia indexing and retrieval. In particular the gamut of techniques from supervised to unsupervised systems will be demonstrated.

© 2004 Elsevier Inc. All rights reserved.

Keywords: Factor graphs; Sum product algorithm; Active learning; Hidden Markov models; Dynamic Bayesian networks; Support vector machines

1. Introduction

Analyzing the semantics of multimedia content is essential and for popular utilization of multimedia repositories. Various multimedia applications such as storage and retrieval, transmission, editing, mining, commerce etc., require the availability of semantic metadata along with the content. The MPEG-7 (FCD Information

E-mail address: naphade@us.ibm.com.
Technology, 2001) standard provides a mechanism for describing this metadata. But the challenge is to encode MPEG-7 descriptions automatically or semi-automatically. This is possible if computational multimedia features can be mapped to high-level semantic concepts represented by the media.

This paper analyzes why machine learning is an integral component in a system that aims to map low-level computable multimedia features and high-level semantic concepts. The focus is on statistical techniques for video classification, structuring, summarization, filtering, and browsing. Having established the premise that learning is integral to semantic media analysis this paper then analyzes the challenges that such techniques face. Using the recent TREC Video Corpus as a test-bed, this paper illustrates the various techniques used for modeling semantic concepts, and the context of inter-conceptual correlations. The paper also discusses the role of supervision, the amount and nature of user-interaction, and methods to reduce this supervision without compromising on the performance of concept modeling. The structure of the paper is organized to cover the supervised, unsupervised, and actively supervised statistical media analysis techniques in that order. Section 1.1 starts with a brief review of the state of the art in statistical media analysis and of the new NIST benchmark in multimedia retrieval. Section 2 discusses statistical semantic concept modeling techniques. Section 3 discusses statistical models for context and knowledge. Section 4 discussed unsupervised techniques for discovering events in media content. Finally Section 5 discusses techniques where the system is actively participating in the process of supervision to reduce it significantly.

1.1. Review

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

- Detecting high-level structures like dialogs and scenes, commercials (Fischer et al., 1995) and other such recurring patterns common to multiple domains (Ferman et al., 1999; Nam et al., 1997; Naphade and Huang, 2000, 2002; Wolf, 1997).
- Classifying the genre or the topic of the video clip. Examples include the MoCA system (Fischer et al., 1995), and Kobla et al. (2000), Liu et al. (1998), and Iyengar and Lippman (1998). Such categorization has its roots in similar approach for topical detection from text.
- Domain-dependent analysis like that of sports video [football (Courtney, 1997)], and news video Wactlar et al., n.d. In such systems, the knowledge is built manually by the human expert in the form of rules. Brand et al. (1997) use coupled HMMs to model complex actions in Tai Chi movies.
- Domain-independent analysis for automatic annotation of video. This is the most challenging multimedia analysis problem. This necessitates the detection of semantic concepts in video (Naphade et al., 1998, 2000a,b; Qian et al., 1999).

Similarly the medium of analysis can be used to distinguish between the state of the art techniques.
• Image and image sequence analysis: semantic classification schemes include those for images (Barnard and Forsyth, 2001; Naphade et al., 2000c; Ratan et al., 1999; Vailaya et al., 1998) as well as those for image sequences (Fischer et al., 1995; Naphade et al., 1998; Qian et al., 1999).

• Auditory scene analysis: This 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 (Akutsu et al., 1998; Ellis, 1996; Jang and Hauptmann, 1999; Liu et al., 1998; Naphade and Huang, 2000; Wold et al., 1996; Zhang and Kuo, 2000).

• Audiovisual Analysis: Of the several techniques referred to in this paper, only a few (Clarkson and Pentland, 1999; Fischer et al., 1995; Nakamura and Kanade, 1997; Naphade, 2001; Naphade and Huang, 2001; Naphade et al., 2001b; Rehg et al., 1999) can claim to be truly multimodal. Most techniques, using audiovisual data perform temporal segmentation on one medium and then analyze the other medium as described. Examples include (Nam et al., 1997; Wang et al., 2000), and the Informedia project (Wactlar et al., n.d.) 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 (Nakamura and Kanade, 1997), sports (Kobla et al., 2000; Zhang and Kuo, 2000), meeting videos (Foote et al., 1999) etc. Wang et al. (2000) survey a few techniques for analysis using a similar approach for similar domains. Other techniques for video analysis include the unsupervised clustering of videos (Clarkson and Pentland, 1999). Naphade et al. (2001b) have presented an algorithm to support query by audiovisual content that uses dynamic programming for the alignment problem. 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 (Rehg et al., 1999). Recent work in domain independent semantic video analysis includes Naphade et al. (1998), Naphade and Huang (2001), and Naphade et al. (2001a).

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 the stage at which a human is involved in the system. 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). Learning can alleviate this interaction (Naphade et al., 2000b, 2002c).

1.2. The NIST TREC video benchmark

The National Institute for Standards and Technology organized the TREC Video benchmark in 2001 and 2002 (TREC, 2001). The goal is to encourage research in multimedia information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. This track is devoted to research in automatic segmentation, indexing, and
content-based retrieval of digital video. The benchmark contained the query answering and shot segmentation tasks in 2001 and 2002. In 2002 a novel concept detection task was added to the track and ten benchmark concepts were used for evaluation. The topics in benchmark queries contain not only text but possibly examples (including video, audio, and images) of what is needed. The topics express a wide variety of needs for video clips: of a particular object or class of objects, of an activity/event or class of activities/events, of a particular person, of a kind of landscape, on a particular subject, using a particular camera technique, answering a factual question, etc. The unit of retrieval is a shot. The test data consisted of about 11h of NIST documentary video clips comprising of over 7000 shots in 2001 and 40h of open source videos in 2002. There were 74 queries in 2001 and 25 semantic queries in 2002. An example is “Find all space shuttle liftoffs.”

For the Video TREC 2002 Concept Detection Benchmark NIST provided a data set of 24h of MPEG video for the concept development and later tested the detectors using a 5h test set. NIST defined non-interpolated average precision over 1000 retrieved shots as a measure of retrieval effectiveness. Let \( R \) be the number of true relevant documents in a set of size \( S \); \( L \) the ranked list of documents returned. At any given index \( i \) let \( R_i \) be the number of relevant documents in the top \( i \) documents. Let \( I_j = 1 \) if the \( j \)th document is relevant and 0 otherwise. Assuming \( R < S \), the non-interpolated average precision (AP) is then defined as

\[
\frac{1}{R} \sum_{j=1}^{S} \frac{R_j}{j} \cdot I_j.
\]

2. Statistical methods for semantic concept modeling

The generic framework for modeling semantic concepts from multimedia features (Naphade et al., 1998) includes an annotation system to create a labeled training set, a learning framework for building models and a detection module for ranking unseen content based on detection confidence for the models (which can be interpreted as keywords). The system must account for the uncertainty in the information in multimedia representation. Suitable learning models include generative models (Naphade et al., 2002a) as well as discriminant techniques (Naphade and Smith, 2003a). Positive examples for interesting semantic concepts are usually rare. In this situation it turns out that discriminant classification using support vector machines (Vapnik, 1995) performs better. These concept models were referred to as multijets or probabilistic multimedia objects (see Naphade et al., 1998). A multijet represents a semantic concept in terms of a probabilistic function of the multiple media features. Multijets belong to one of the three categories: objects (car, man, and helicopter), sites representing static concepts like sky, mountain, outdoor, cityscape, etc., or events representing a chain of temporal activities (explosion, man-walking, and ball-game). This categorization of semantic concepts is not rigorous, but helps in choosing models for the three concept classes. Sites are modeled with static pattern
modeling 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.

2.1. Supervised concept learning

The supervised concept learning approach expects a set of patterns and annotations or labels for the semantic concepts. 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. Subsequently we can use indirect modeling techniques such as density modeling or direct techniques such as modeling the decision boundary.

In the density modeling approach we assume that under each hypothesis, the features are generated by the conditional probability density functions \( P_i(X), i \in \{0, 1\} \). For distinct instances of all multijets, further assume, that these features are independent identically distributed random variables drawn from known probability distributions, with unknown deterministic parameters. Application of density modeling and hypothesis testing for semantic concept modeling include Naphade et al. (1998, 2002c), and Vailaya et al. (1998). The other popular approach is to directly learn the decision boundary and examples include Support vector machines (SVM), neural nets, decision trees etc., classifiers (Vapnik, 1995). SVMs project the original feature dimension into a high-dimensional space using nonlinear kernel functions and attempt to find that linear separating hyperplane in the higher dimensional space, which maximizes generalization capability.

For experiments reported here the models were built using features extracted from key-frames.

Fig. 1 shows the feature and parameter selection process incorporated in the learning framework for optimal model selection and is described below.

---

Fig. 1. SVM learning: optimizing over multiple possible feature combinations and model parameters.
We extract features for color, texture, shape, structure etc. We then perform feature type selection. We simply concatenate one or more of these feature types (appropriately normalized). Different combinations of features (i.e., only color, color and texture, color and shape etc.) can then be used to construct models and the validation set is used to choose the optimal combination. This is feature selection at the coarse level of feature types. Results of this feature type combination selection and early fusion are presented in Section 2.2.3. The performance of SVM classifiers can vary significantly with variation in parameters of the models. Choice of the kernels and their parameters is therefore crucial. To reduce sensitivity to these design choices, we experiment with different kernels and for each kernel we build models for several combinations of the parameters. Radial basis function (RBF) kernels usually perform better than other kernels. In our experiments we built models for different values of the RBF parameter \( \gamma \) (variance), relative significance of positive vs. negative examples \( j \) (necessitated also by the imbalance in the number of positive vs. negative training samples) and trade-off between training error and margin \( c \). Using the validation set we then performed a grid search for the combination that resulted in highest average precision.

2.2. Experimental setup

2.2.1. Lexicon

We created a lexicon with more than hundred semantic concepts for describing events, sites, and objects (Naphade et al., 1998). However only 34 concepts had support of more than 20 shots in the training set and were modeled:

- Scenes: Outdoors, Indoors, Landscape, Cityscape, Sky, Greenery, Water-body, Beach, Mountain, Land, Farm Setting, Farm Field, Household Setting, Factory Setting, and Office Setting.
- Objects: Face, Person, People, Road, Building, Transportation Vehicle, Car, Train, Tractor, Airplane, Boat, Tree, Flowers, Fire/smoke, Animal, Text Overlay, Chicken, Cloud, and Household Appliances.

2.2.2. Feature extraction

After performing shot boundary detection and key-frame extraction (Srinivasan et al., 2000) each key-frame was analyzed to detect the 5 largest regions described by their bounding boxes.

The system then extracts the following low-level visual features at the frame-level or global level as well as the region level for the entire frame as well as each of the regions in the key-frames.

*Color histogram* (72): 72-bin YCbCr color space \( (8 \times 3 \times 3) \).

*Color Correlogram* (72): Single-banded auto-correlogram coefficients extracted for 8 radii depths in a 72-bin YCbCr color space.

*Edge Orientation Histogram* (32): Using a Sobel filtered image and quantized to 8 angles and 4 magnitudes.
Co-occurrence Texture (48): Based on entropy, energy, contrast, and homogeneity features extracted from gray-level co-occurrence matrices at 24 orientations (Naphade et al., 2002c).

Moment Invariants (6): Based on Dudani’s moment invariants (Naphade et al., 2002c) for shape description.

Normalized Bounding Box Shape (2): The width and the height of the bounding box normalized by that of the image.

2.2.3. Validation set performance

Fig. 2 demonstrates the importance of parameter selection of the SVM models. Exhaustive modeling for different parameter combinations and use of validation set for selection helps significantly in minimizing sensitivity of the model performance as seen from the range of AP from 0.15 to 0.53 in this case. Fig. 2 in particular shows the precision recall curves for 12 parameter combinations of $c$, $j$, and $c$ of the RBF kernel for the co-occurrence feature type. In this case it is clear that $j = 4$ is a bad choice irrespective of the other parameters. Fig. 3 displays bar plots for all 34 semantic concepts. We compare average precision for each concept with the ratio of positive training samples to the total number of training samples for that concept. The number of positive training samples vary from 20 (Beach with AP 0.17) to 2809 (Outdoors with AP 0.59).

2.2.4. 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 (Rabiner, 1989) and its several variants for fusing multiple modalities like the coupled HMM (Brand et al., 1997), factorial HMM (Ghahramani and Jordan, 1997), etc. Fusion of multimodal feature streams (especially audio and visual feature streams) has been applied to problems like bimodal speech (Chen and Rao, 1998), summarization of video (Nakamura and Kanade, 1997), query by audiovisual content (Naphade...
et al., 2001b), and event detection in movies (Naphade et al., 1998). 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 (Brand et al., 1997), the factorial HMM (Ghahramani and Jordan, 1997), the duration dependent input output Markov models (DDIOMM) (Naphade et al., 2001a) 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.

3. Statistical methods for representation of knowledge and context

Semantics is meaningful only in context and not in vacuum. This fact has been used in information retrieval in the text domain by WordNet (Miller, 1995). WordNet is a text-retrieval system built manually over several years that encodes relation-
ships of different kinds between the words supported in the system. This information may be domain dependent. The constraints may be represented by rules heuristics or association. Semantic concepts occur within a well-defined context. Thus the presence of certain semantic concepts suggests a high possibility of other concepts. Similarly, some concepts are less likely to occur in the presence of others. The detection of sky and greenery boosts the chances of detecting a Landscape, and reduces the chances of detecting Indoors. It might also be possible to detect some concepts and infer more complex concepts based on their relation with the detected ones. Detection of human speech in the audio stream and a face in the video stream may lead to the inference of human talking. To integrate all the concepts and model their interaction, we proposed the network of multijets (Naphade et al., 1998) which we termed as a multinet (Naphade et al., 2002c). A conceptual figure of a multinet is shown in Fig. 4 with positive (negative) signs indicating positive (negative) interaction.

Discriminant training has been shown to improve detection performance of the concept models (multijets) (Naphade and Smith, 2003b). On the other hand, the graphical multinet that we proposed (Naphade et al., 2002c) for a small number of concepts demonstrated performance improvement over generatively trained multijets. In this paper, multijet models built using discriminant training are integrated within a graphical multinet that models joint probability mass functions of semantic concepts. We also use a probabilistic factor graph framework, which models the interaction between concepts within each video shot as well as across the video shots within each video clip. Factor graphs provide an elegant framework to represent the stochastic relationship between concepts, while the sum-product algorithm provides an efficient tool to perform learning and inference in factor graphs. Using exact as well as approximate inference (through loopy probability propagation) we show that

Fig. 4. A conceptual multinet showing relations between 12 visual concepts. Nodes in the network represent semantic concepts. Edges represent relationships. Symbols on the edges can be interpreted as indicative of the nature of interaction between concepts linked by the edge.
there is significant improvement in the detection performance. Using the TREC Video 2002 Benchmark Corpus and some of the benchmark concepts (Adams et al., 2002), we show that explicit modeling of concept relationships and the use of this model for enforcing inter-conceptual and temporal relationships leads to an improvement in detection performance. We also provide efficient approximations to the context models that are scalable in terms of number of concepts being modeled. We show that these approximations retain the significant improvement in detection performance.

3.1. A factor graphical context model

To model the interaction between multijets in a multinet, we proposed a factor graph (Kschischang et al., 2001; Naphade et al., 2002c) framework. Factor graphs subsume graphical models like Bayesian nets and Markov random fields and have been successfully applied in the area of channel error correction coding (Kschischang et al., 2001) and specifically, iterative decoding. Let \( x = \{x_1, x_2, \ldots, x_n\} \) be a vector of variables. A factor graph visualizes the factorization of a global function \( f(x) \). Let \( f(x) \) factor as

\[
f(x) = \prod_{i=1}^{m} f_i(x^{(i)}),
\]

where \( x^{(i)} \) is the set of variables 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 \( i \)th node in \( V_f \) is connected to the \( j \)th node in \( V_v \) iff \( f_i \) is a function of \( x_j \). For details please see Kschischang et al. (2001). We will use rectangular blocks to represent function nodes and circular nodes to represent variable nodes. (See Fig. 5.)

Many signal processing and learning problems are formulated as optimizing a global function \( f(x) \) marginalized for a subset of its arguments. The sum-product algorithm allows us to perform this efficiently, though in most cases only approximately. It works by computing messages at the nodes using a simple rule and then passing the messages between nodes according to a reasonable schedule. For details on message computations see Kschischang et al. (2001). 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 (Kschischang et al., 2001) leading to loopy probability propagation. Because rela-

![Fig. 5. An example of function factorization.](f5.png)
tions between semantic concepts are complicated and in general contain numerous cycles (e.g., see Fig. 4) this provides the ideal framework for modeling context.

3.2. Modeling inter-conceptual context in a factor graph

We now describe a shot-level factor graph to model the probabilistic relations between various frame-level semantic features $F_i$ obtained by using the distance of the test set examples from the separating hyperplane as a measure of confidence. To capture the co-occurrence relationship between the twelve semantic concepts at the frame-level, we define a function node which is connected to the twelve variable nodes representing the concepts as shown in Fig. 6. Since we will present results using 12 concepts, the multinet in this paper depict 12 variable nodes.

The function node depicted by the rectangular box at the top represents the joint mass function $P(F_1, \ldots, F_N)$. The function nodes below the variable nodes provide the individual SVM-based multiject $P(F_i = 1 \mid X), P(F_i = 0 \mid X)$, where $F_i$ is the $i^{th}$ concept and $X$ is the observation (features). These are then propagated to the function node. At the function node the messages are multiplied by the joint mass function estimated from 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 twelve concepts.

In general, the distribution at the function node in Fig. 6 is exponential in the number of concepts ($N$) and the computational cost may increase quickly. To alleviate this we can enforce a factorization of the function in Fig. 6 as a product of several local functions where each local function accounts for co-occurrence of two variables only. This modification is shown in Fig. 7.

Each function at the top in Fig. 7 represents the joint probability mass of those two variables that are its arguments (and there are $C_N^2$ such functions) thus reducing the complexity from exponential to polynomial in the number of concepts. For 12 concepts, there are 66 such local functions and the global function is approximated as a product of these 66 local functions. The factor graph is no longer a tree and exact inference becomes hard as the number of loops grows. We then apply iterative techniques based on the sum-product algorithm to overcome this.

---

Fig. 6. A factor graph multinet with 12 concepts represented by variable nodes, 12 concept detectors denoted by the function nodes at the bottom and one global context function denoted by the function node at the top. In this paper the global context function used is the joint mass function $P(F_1, \ldots, F_N)$. 
In addition to inter-conceptual relationships, we can also incorporate temporal dependencies. This can be done by replicating the slice of factor graph in Figs. 6 or 7 as many times as the number of frames within a single video shot and by introducing a first order Markov chain for each concept.

Figs. 8 and 9 show two consecutive time slices and extend the models in Figs. 6 and 7, respectively. The horizontal links in Figs. 6 and 7 connect the multinet instances in consecutive shots.

If each concept represented by the variable node within each shot is assumed to be binary then a state vector of $N$ such binary concepts can take $2^N$ possible configurations and thus the transition matrix $A$ is $2^N \times 2^N$ dimensional. However we make an independence assumption across time that results in a simplified transition matrix.

![Fig. 8. Replicating the multinet in Fig. 6 for each shot (key-frame) and introducing temporal dependencies between the value of each concept in consecutive shots. Within each state the unfactored global multinet of Fig. 6 is used. The temporal transitions are modeled by a stochastic transition matrix $A$. The function nodes shown in between the shots use this matrix $A$ to determine $P(s_i \mid s_{i-1})$. $\pi_0$ is the prior on the state.](image)
We assume that each concept's temporal context can be modeled independent of the other concepts in the past and the future. Since we are already accounting for within-concept dependencies using the multinet at each time instant or shot, this assumption is fair. This results in the simplification of the temporal context estimation. Instead of estimating the entire $A$ matrix, we now estimate $N \times 2$ transition matrices. Each variable node for each concept in a time slice can thus be assumed to be connected to the corresponding variable node in the next time slice through a function modeling the transition probability $A_c, c \in 1, \ldots, N$. 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.

3.4. TREC Video 2002 Corpus validation set performance

For experiments reported here we will confine to those 12 concepts that had a support of 300 or more examples in the training set:

- Scenes: Outdoors, Indoors, Landscape, Cityscape, Sky, and Greenery.
- Objects: Face, Person, People, Road, Transportation Vehicle, and Tree.

These 12 multiject models were trained using SVM classifiers as described in Section 2.2.3 (Naphade and Smith, 2003a) although the factor graph multinet itself is agnostic to the individual concept detection model be it a generative model like a GMM or an HMM (Naphade et al., 2002c) or a discriminant model like the SVM. This is the baseline for comparison with the context enforced detection. The detection confidence is then modified using the multinet to enforce context. We evaluate performance of the globally connected multinet of Fig. 6 model as well as the
factored model of Fig. 7. The comparison using the average precision measure (Eq. (1)) is shown in Table 1.

The improvement in mean average precision using the global unfactored multinet is 17%. The approximation of the global multinet using the factored version also improves the average precision by 14%. Across all concepts the maximum improvement is as much as 117% in the case of Road. Also no concept suffers any deterioration in performance. Using the temporal context in conjunction with the inter-conceptual context improves mean average precision by 20.44% for the unfactored multinet and 20.35% for the factored approximation. Thus the factored approximation results in performance almost identical to the global function. Importantly, this improvement is achieved without using any additional training data or annotation.

Fig. 10 shows a set of precision recall curves with and without the various proposed context models, for the concept Road. The average precision as well as the curves show that detection is significantly improved by accounting for context. Inter-conceptual context improves detection and accounting for temporal context in addition improves detection further. Also the factored context models result in performance very similar to the unfactored context models in the static as well as dynamic case. Inter-conceptual and temporal context modeling and enforcement thus improves detection performance for Road by 117%.

4. Unsupervised detection of recurring temporal patterns and structure

Statistical models like the hidden Markov models (HMM) and its various extensions have been used to impose structure on videos (Ferman et al., 1999; Liu and Kender, 2000; Liu et al., 1998; Wolf, 1997), improve the detection of shots and
scenes (Nakamura and Kanade, 1997; Nam et al., 1997; Srinivasan et al., 2000). Rules of production have been studied and applied (Kender and Yeo, 1998; Sundaram and Chang, 2000) to detect recurring patterns such as dialogues classify TV programs into types of sports, news, weather etc. Clarkson and Pentland (1999) analyze the audio-visual signals from a wearable camera and a microphone. Naphade and Huang (2002) proposed a hierarchical hidden Markov model where non-ergodic hidden Markov models are embedded within a long-term HMM 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. The novel aspect of this algorithm is its ability to account for short-term as well as longer term continuity and its ability in detecting recurring patterns automatically without supervision.

Our approach is motivated by the unsupervised clustering algorithm of Poritz (1982) that was extended to large vocabulary speech recognition by Levinson (1989). The idea is to use an ergodic hidden Markov model whose state transitions are illustrated in Fig. 11 to cluster the signal into as many stationary patterns as the number of states in the HMM. However an HMM state typically handles stationarity of short-term statistics in the signal. We want to cluster temporal patterns that span multiple states. To give a textual analogy, if the traditional Poritz model can detect recurring patterns and clusters in alphabets we are interested in recurring strings of such alphabets. Here we present a natural extension to overcome this challenge. The architecture is shown in Fig. 11B. The assumption is that the time series is produced by the generative model of Fig. 11B. Just as in HMMs the observation at any given time is dependent only on an unobserved state and the current unobserved state dependent on the state in the previous time instant. The framework can be
thought of as a number of non-ergodic hidden Markov models embedded within a hierarchical ergodic hidden Markov model. Each state in Fig. 11A is thus replaced by a non-ergodic HMM as in Fig. 11B with states labeled $D$ acting as dummy states without emitting any observations. The model can thus significantly improve the modeling of the long-term statistics of a piece-wise stationary signals. Interestingly the signal can be analyzed at a coarser temporal granularity that is offered by the detection of each branch in Fig. 11B. The model is trained using the expectation maximization (EM) algorithm (Dempster et al., 1977). We applied this algorithm to two different domains. In both domains we extract commonly used audiovisual features. We extracted popular audiovisual features such as linearized HSV histogram (H: 6 bins, S: 6 bins, V: 12 bins), a 24 bin edge direction (Naphade et al., 2001b) histogram to represent the structure in each frame within the image sequence, and 15 mel frequency cepstral coefficients, 15 delta coefficients and 2 energy coefficients extracted by using a 33ms window width for representing audio characteristics. The audio processing window was deliberately chosen to equal the duration of 1 visual frame for simplicity in synchronization.

We chose ten video clips from an action thriller that has several examples of interesting audiovisual events such as explosion. For the audio feature stream we applied the proposed model with 6 possible non-ergodic branches each with 3 hidden states. Each state emits observations using a Gaussian mixture model (GMM) with 5 mixture components. The GMM in each state of the proposed model was assumed to have a diagonal covariance matrix. For the visual stream we experimented with distinct models for color distribution, and structure as well as a model for the combined visual feature set. In each case we used 5 non-ergodic branches. One of the non-ergodic HMM chains in the model picked up all the auditory events of explosion and crash like sounds. Another picked up male speech while a third picked up a recurring musical theme that is played in different styles in the movie. The visual
models also similarly converged to events such as the evolution of a fire and smoke, outdoor locales, indoor locales with people and faces etc. Fig. 12 shows the different shots for which the same chain was active. All shots correspond to the event explosion.

5. Semi-supervised methods for minimizing user interaction

From the application viewpoint, user interaction comes into picture in video indexing and retrieval systems in one of the two scenarios. In the first, a user annotates the video. In the second, the user changes system parameters to enhance performance. 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. For this we need to learn to represent the user’s inputs as well as the user’s preferences. Annotation can be assisted through intelligent modeling of the user inputs and through propagation and sample selection procedures. By exploiting the inherent clusters in the data space we can cut down on the number of training samples needed significantly.

Active learning strategies can be broadly classified into three different categories. Uncertainty sampling, Query by Committee and Adaptive re-sampling. A boosting-like technique that adaptively resample data biased towards the misclassified points in the training set and then combine the predictions of several classifiers is also used (Iyengar et al., 2000). Even among the uncertainty sampling methods a variety of classifiers and measures of degree of uncertainty of classification have been proposed. Two specific classifiers that have been widely used for this purpose are the Support Vector Machine (SVM) classifier and gaussian Mixture Model based classification scheme. For SVM classifier the distance of an unlabeled data-point from the separating hyperplane in the high-dimensional feature space could be taken as a measure of uncertainty (Tong and Chang, 2001) (alternatively, a measure of confidence in classification) of the data-point. For the gaussian mixture model based classifier the likelihood of the new data-point given the current mixture model can be...
used as measure of this confidence. Our approach will be more akin to the SVM approach (Zhang and Oles, 2000) but we are interested in applying it for persistent semantic annotation.

5.1. Experiments with SVM based active learning

We now demonstrate the effectiveness of our support vector machine based active learning algorithm on TREC Video 2001 corpus. We use polynomial kernel machines\(^1\) We report results using the TREC Video 2001 database for annotation of on indoor-outdoor images. There are 9045 examples in the database to be annotated. We begin with a warm-up set of 1% of annotated data.\(^2\) The initial SVM classifier is built using this warm-up set. New examples are then presented to the classifier in steps. Each unseen example is classified by the SVM classifier and the confidence in classification is taken to be inversely proportional to the distance of the new feature from the separating hyperplane. If this distance is less than a specified threshold then we select the new sample to be included in the training set. We evaluate three different selection strategies for choosing the next sample to be annotated: The sample closest to the current hyperplane (type-I); The sample closest to the hyperplane that has been previously classified as a negative class (type-II); and The sample closest to the hyperplane where distance from the hyperplane is rescaled using the ratio points classified negatively to points classified positively (type-III). The type-II selection prefers false-positives over false-negatives. The type-III selection chooses on average more samples from the infrequent class. The SVM classifier is retrained after every decision to include a new example based on the selection process in the training set. Iterative updates of the classifier can proceed in this manner until a desirable performance level is reached. In our experiments we stop when 10% of the total samples have been annotated. To evaluate the classifiers obtained at the end of 10% annotation, we create a baseline as well as near-ideal situation. The baseline (the lowermost dotted curve) uses passive annotation with 10% training. The near-ideal situation chooses passive annotation with 90% of the samples annotated (uppermost continuous curve). In Fig. 13 we compare the precision recall curves for of the active classifiers to those of the 2 passive classifiers. The other 3 curves illustrate the performance of the 3 selection schemes with 10% actively chosen samples used for incremental training. It is remarkable that with all three active selection strategies, active learning with only 10% data shows performance almost as good as passive training with 90% data and much better than passive training with 10% data. Other ways of reducing annotation effort include the use of unlabeled samples along with the labeled samples to improve the performance of automatic propagation of annotations (Naphade et al., 2000c).

\(^1\) The SVMlight software (Joachims, 1999) is used for simulations.

\(^2\) The warm-up set can be selected randomly. It can also be selected by unsupervised clustering (Naphade et al., 2000c).
6. Concluding remarks and future directions

The influence of statistical learning in semantic video analysis for search, segmentation and summarization continues to increase. The learning component attempts to bridge the gap between low-level media features and representations that can be computed and the high-level semantic labels. 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 with minimal user input.

Acknowledgments

Thomas S. Huang at the University of Illinois at Urbana-Champaign and John R. Smith at the IBM Thomas J. Watson Research Center for their comments and inputs. The IBM TREC Video Team for annotation, features, and data set processing. Ching-Yung Lin for bounding boxes.

References


