EURASIP Journal on Applied Signal Processing 2003:2, 1–16

Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues

W. H. Adams
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
Email: whadams@us.ibm.com

Giridharan Iyengar
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
Email: giyengar@us.ibm.com

Ching-Yung Lin
IBM Thomas J. Watson Research Center, Hawthorne, NY 10532, USA
Email: chingyung@us.ibm.com

Milind Ramesh Naphade
IBM Thomas J. Watson Research Center, Hawthorne, NY 10532, USA
Email: naphade@us.ibm.com

Chalapathy Neti
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
Email: chalapathy_Neti@us.ibm.com

Harriet J. Nock
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
Email: hnock@us.ibm.com

John R. Smith
IBM Thomas J. Watson Research Center, Hawthorne, NY 10532, USA
Email: jrsmith@watson.ibm.com

Received 2 April 2002 and in revised form 15 November 2002

We present a learning-based approach to the semantic indexing of multimedia content using cues derived from audio, visual, and text features. We approach the problem by developing a set of statistical models for a predefined lexicon. Novel concepts are then mapped in terms of the concepts in the lexicon. To achieve robust detection of concepts, we exploit features from multiple modalities, namely, audio, video, and text. Concept representations are modeled using Gaussian mixtures models (GMM), hidden Markov models (HMM), and support vector machines (SVM). Models such as Bayesian networks and SVMs are used in a late-fusion approach to model concepts that are not explicitly modeled in terms of features. Our experiments indicate promise in the proposed classification and fusion methodologies: our proposed fusion scheme achieves more than 10% relative improvement over the best unimodal concept detector.

Keywords and phrases: query by keywords, multimodal information fusion, statistical modeling of multimedia, video indexing and retrieval, SVM, GMM, HMM, spoken document retrieval, video event detection, video TREC.

1. INTRODUCTION

Large digital video libraries require tools for representing, searching, and retrieving content. One possibility is the query-by-example (QBE) approach, in which users provide (usually visual) examples of the content they seek. However, such schemes have some obvious limitations, and since most users wish to search in terms of semantic-concepts rather

than by visual content [1], work in the video retrieval area has begun to shift from QBE to query-by-keyword (QBK) approaches, which allow the users to search by specifying their query in terms of a limited vocabulary of semantic-concepts. This paper presents an overview of an ongoing IBM project which is developing a trainable QBK system for the labeling and retrieval of generic multimedia semantic-concepts in video; it will focus, in particular, upon the detection of semantic-concepts using information cues from multiple modalities (audio, video, speech, and potentially video-text).[1]

1.1. Related work

Query using keywords representing semantic-concepts has motivated recent research in semantic media indexing [2, 3, 4, 5, 6, 7, 8]. Recent attempts to introduce semantics in the structuring and classification of videos includes [9, 10, 11, 12].

Naphade et al. [2] present a novel probabilistic framework for semantic video indexing by learning probabilistic multimedia representations of semantic events to represent keywords and key concepts. Chang et al. [3] use a library of examples approach, which they call semantic visual templates. Zhang and Kuo [5] describe a rule-based system for indexing basketball videos by detecting semantics in audio. Ellis [6] presents a framework for detecting sources of sounds in audio using such cues as onset and offset. Casey [8] proposes a hidden-Markov-model (HMM) framework for generalized sound recognition. Scheirer and Slaney [13] investigate a variety of statistical models including Gaussian mixtures models (GMMs), maximum a posteriori classifiers, and nearest neighbors for classification of speech and music sounds.

There has been also work in detecting the semantic structure (emphasizing the temporal aspects of it) in video. Wolf [9] use HMMs to parse video. Ferman and Tekalp [11] attempt to model semantic structures such as *dialogues* in video. Iyengar and Lippman [10] present work on genre classification by modeling the temporal characteristics of such videos using an HMM. Adams et al. [14] propose using *tempo* to characterize motion pictures. They suggest a computational model for extracting tempo information from a video sequence and demonstrate the usefulness of this feature for the structuring of video content.

In prior work, the emphasis has been on the extraction of semantics from individual modalities, in some instances, using audio and visual modalities. We are not aware of any work that combines audio and visual content analysis with textual information retrieval for semantic modeling of multimedia content. Our work combines content analysis with information retrieval in a unified setting for the semantic labeling of multimedia content. In addition, we propose a novel approach for representing semantic-concepts using a basis of other semantic-concepts and propose a novel discriminant framework to fuse the different modalities.

[1] *Audio* here refers to the nonspeech content of the sound track.

1.2. Our approach

We approach semantic labeling as a machine learning problem. We begin by assuming the a priori definition of a set of *atomic* semantic-concepts (objects, scenes, and events) which is assumed to be broad enough to cover the semantic query space of interest. By atomic semantic-concepts, we mean concepts such as sky, music, water, speech, and so forth, which cannot be decomposed or represented straightforwardly in terms of other concepts. Concepts that can be described in terms of other concepts are then defined as *high-level* concepts. Clearly, the definition of high-level concepts depends, to some extent, on the variety of atomic concepts defined; this distinction is being made for practical purposes. We note that these concepts are defined independently of the modality in which they are naturally expressed (i.e., an atomic concept can be multimodal and a high-level concept can be unimodal etc.).

The set of atomic concepts are annotated manually in audio, speech, and/or video within a set of "training" videos. Examples of concepts occurring in audio include rocket engine explosion, music, and speech, for video, an outdoor scene, a rocket object, fire/smoke, sky, and faces. The annotated training data is then used to develop explicit statistical models of these atomic concepts; each such model can then be used to automatically label occurrences of the corresponding concept in new videos. However, semantic-concepts of interest to users are often at the high-level. Examples of these high-level concepts are typically sparse. Thus, rather than constructing models for each of these high-level concepts directly as for the atomic concepts, more complicated statistical models are constructed that combine information from existing atomic (or even higher-level) models as well as the information in the manually labeled training data. As with atomic concepts, the resulting high-level semantic models are then used to label new videos.

There are several challenges to be overcome in such a system. Firstly, low-level features appropriate for labeling atomic concepts must be identified (different features may be appropriate for different concepts) and appropriate scheme(s) for modeling these features are to be selected. The paucity of examples for many concepts will be an important factor in the choice of a modeling scheme. In addition, we need techniques for segmenting objects automatically from video. This paper assumes that segmented regions are available both for training and testing. However, we have investigated automated segmentation from video as part of our ongoing work. Secondly, high-level concepts must be linked to the presence (or absence) of other concepts (either within a modality or across) and statistical models for combining these concept models into a high-level model must be chosen. Thirdly, cutting across these levels, information from multiple modalities must be integrated or *fused*. Fusion could occur at various levels: low-level features, within atomic concept models, or by combining several atomic-concept models within a multimodal high-level concept models. In this paper, we focus on the modeling of atomic concepts and on the representation of high-level concepts. We use a standard

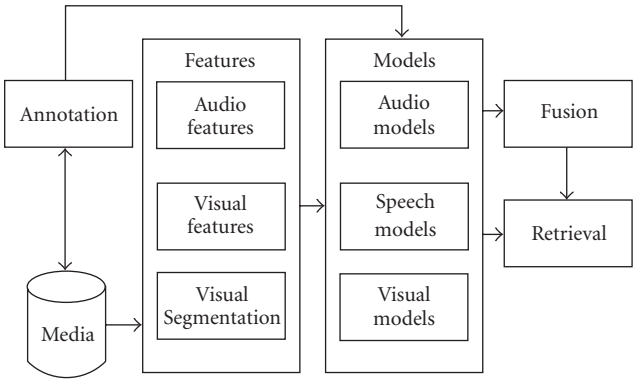


Figure 1: Diagram of semantic-concept analysis system.

set of low-level features that is well established in literature.

The rest of the paper is described below. Section 2 presents a detailed overview of the proposed semantic-content analysis system. Sections 2.1 and 2.2 describe the concept lexicon and the annotation process. Section 2.3 describes schemes for semantic-concept modeling. Sections 2.4, 2.5, and 2.6 detail the single-modality concept-modeling schemes used. Section 2.7 then describes schemes for concept retrieval which integrate cues from all of the modalities. Section 3 evaluates these techniques using the NIST 2001 Video TREC corpus [15]. The paper ends with conclusions and outlines possible future work.

2. SEMANTIC-CONTENT ANALYSIS SYSTEM

The proposed IBM system for semantic-content analysis and retrieval comprises three components:

(i) tools for defining a *lexicon* of semantic-concepts and *annotating examples* of those concepts within a set of training videos;

(ii) schemes for automatically *learning the representations* of semantic-concepts in the lexicon based on the labeled examples;

(iii) tools supporting *data retrieval* using the (defined) semantic-concepts.

As a starting point, our unit of semantic labeling and retrieval is a camera shot. Future work will address whether this is the most effective unit for semantic-concept labeling. The overall framework is illustrated in Figure 1.

2.1. Lexicon of semantic-concepts

The lexicon of semantic-concepts defines the working set of intermediate- and high-level concepts, covering events, scenes, and objects. These concepts are defined independently of the modality in which their cues occur: whilst some are naturally expressed in one modality over the other (e.g., *music* is an audio concept, whereas *sky* is a visual concept), others require annotation across modalities, for example, a person talking versus, a person singing. The lexicon is, in principle, extendable by users.

While it is difficult to impose a strict hierarchy on semantic-concepts, some of them may be defined in terms of feature representations while others may have to be defined in terms of a set of such concepts themselves. For example, semantic-concepts such as sky, music, water, speech, and so forth, can be represented directly in terms of media feature representations. There are other concepts that may only be inferred partially from other detected concepts and partially from feature representations. For example, the semantic-concept *parade* may need to be described in terms of a collection of people, a particular type of accompanying music, and a particular context in which the video clip may be interpreted as a parade.

2.2. Annotating a corpus

Manually labeled training data is required in order to learn the representations of each concept in the lexicon. We have built tools that allow users to annotate video sequences with concepts from the lexicon. Annotation of visual data is performed at shot level; since concepts of objects (e.g., rockets and cars) may occupy only a region within a shot, tools also allow users to associate object labels with an individual region in a key-frame image by specifying *manual bounding boxes (MBB)*. Annotation of audio data is performed by specifying time spans over which each audio concept (such as speech) occurs. Speech segments are then manually transcribed. Multimodal annotation follows with synchronized playback of audio and video during the annotation process. This permits the annotator to use the video to potentially disambiguate audio events and vice versa. Figure 2 shows the multimodal annotation interface. See also Marc Davis' Media Streams [16] for a video annotation interface. Media Streams presents a lexicon of semantic-concepts in terms of a well-designed set of icons. Media Streams allows for the creation of novel semantic-concepts by allowing users to create compound icons from the lexicon. Media Streams does not, however, allow the annotator to provide explicit object boundaries, audio-segment boundaries, and so forth. Also, Media Streams does not differentiate between audio, visual, and multimodal concepts, as in our annotation interface. In addition, the lexicon used by our annotation interface is an XML document that can be edited from within the tool or can be easily edited/changed with any XML editor. We envision the user switching to appropriate lexicons for different tasks and domains.

2.3. Learning semantic-concepts from features

Mapping low-level features to semantics is a challenging problem. This is further complicated by the paucity of training examples. Given the labeled training data, useful features must be extracted and used to construct a representation of each atomic concept. For the purposes of this paper, human knowledge is used to determine the type of features (e.g., color histograms, motion vectors, pitch trajectories, spectral features, and pertinent words) that are appropriate for each

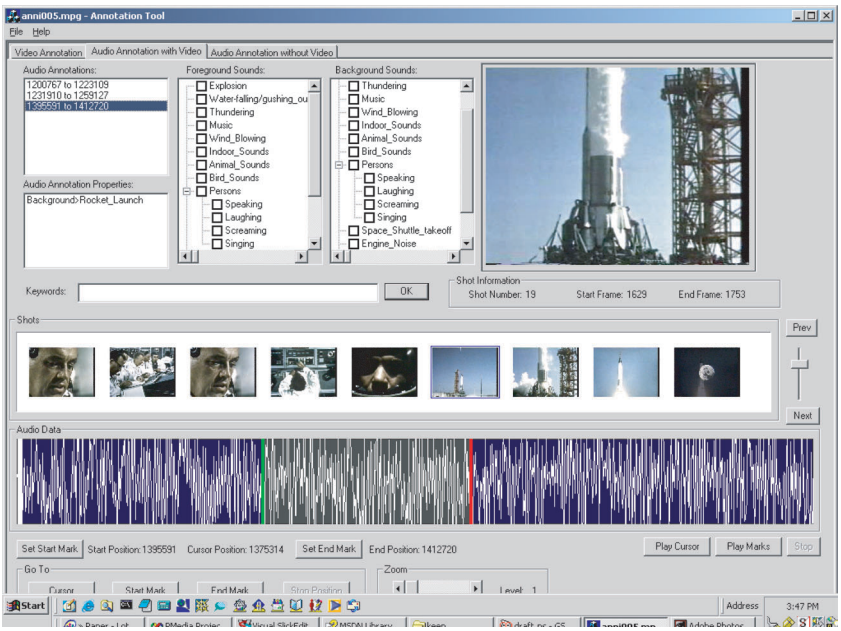


Figure 2: A multimodal annotation interface.

concept; future work will automate this feature selection step. In this paper, atomic concepts are modeled using features from a single modality and the integration of cues from multiple modalities occurs only within models of high-level concepts (a *late integration* approach); use of earlier integration schemes and multimodal models for atomic concepts will be addressed in future work. For instance, Neti et al. [17] have explored several early fusion techniques such as discriminant feature fusion, HMM-based early fusion, and so on, in the context of audiovisual speech recognition. Iyengar and Neti [18] presented an early fusion approach for joint audiovisual speaker change detection. In this paper, the focus is on the joint analysis of audio, visual, and textual modalites for the semantic modeling of video. We employ a late-fusion approach for combining modalities. Since the unit of retrieval is a video shot, our effort has been to focus on fusion at the shot level. However, for some concepts such as monologues, it may be appropriate to focus on the intrashot fusion of modalites.

We now introduce the two main modeling approaches investigated in this paper: probabilistic modeling of semantic-concepts and events using models such as GMMs, HMMs, and Bayesian networks and discriminant approaches such as support vector machines (SVMs).

2.3.1 Probabilistic modeling for semantic classification

In the simplest form, we model a semantic-concept as a class conditional probability density function over a feature space. Given a set of semantic-concepts and a feature observation, we choose the label as that class conditional density which results in the maximum likelihood of the observed feature. In practice, the true class conditional densities are not available, so assumptions must be made as to their form and their parameters estimated using training data. Common choices are GMMs for independent observation vectors and HMMs for time series data.

A GMM [19] defines a probability density function of an n-dimensional observation vector $\mathbf{x}$ given a model M,

$$P(\mathbf{x}|M) = \sum_i \pi_i \frac{1}{(2\pi)^{n/2}\,|\Sigma_i|^{1/2}} e^{-(1/2)(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)}, \quad (1)$$

where μ_i is an n-dimensional vector, Σ_i is an $n \times n$ matrix, and π_i is the mixing weight for the ith gaussian.

An HMM [20] allows us to model a sequence of observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ as having been generated by an unobserved state sequence $s_1, \ldots, s_n$ with a unique starting state s_0, giving the probability of the model M generating the output sequence as

$$P(\mathbf{x}_1, \ldots, \mathbf{x}_n|M) = \sum_{s_1,\ldots,s_n} \prod_{i=1}^{n} p(s_i|s_{i-1})\, q(\mathbf{x}_i|s_{i-1}, s_i), \quad (2)$$

where the probability $q(\mathbf{x}_i|s_{i-1}, s_i)$ can be modeled using a GMM (1), for instance, and $p(s_i|_{i-1})$ are the state transition probabilities. We do maximum-likelihood estimation of both the GMMs and the HMMs using the expectation maximization (EM) algorithm [21].

2.3.2 Discriminant techniques: support vector machines

The reliable estimation of class conditional parameters in the previous section requires large amounts of training data for each class, but for many semantic-concepts of interest, this

may not be available; in addition, the forms assumed for class conditional distributions may not be the most appropriate. Use of a more discriminant learning approach requiring fewer parameters and assumptions may yield better results for this application; SVMs with radial basis function kernels [22] are one possibility.

An SVM tries to find a best-fitting hyperplane that maximizes the generalization capability while minimizing misclassification errors. Assume that we have a set of training samples $(\mathbf{x_1}, \ldots, \mathbf{x_n})$ and their corresponding labels $(y_1, \ldots, y_n)$ where $y_i \in \{-1, 1\}$, then SVMs map the samples to a higher-dimensional space using a predefined nonlinear mapping $\Phi(\mathbf{x})$ and solve a minimization problem in this high-dimensional space that finds a suitable linear hyperplane separating the two classes $(\mathbf{w} \cdot \Phi(\mathbf{x_i}) + b)$, subject to minimizing the misclassification cost,

$$\begin{aligned} \Phi(\mathbf{x_i}) \cdot \mathbf{w} + b &\geq 1 - \epsilon_i \quad \forall\, y_i = 1 \\ \Phi(\mathbf{x_i}) \cdot \mathbf{w} + b &\leq \epsilon_i - 1 \quad \forall\, y_i = -1 \\ \epsilon_i &\geq 0 \quad \forall\, i, \end{aligned} \tag{3}$$

where ϵ_i is a scalar value. If $\mathbf{x_i}$ is to be misclassified, we must have $\epsilon_i > 1$ and hence the number of errors is less than $\sum_i \epsilon_i$. If we add a penalty for misclassifying training samples, it can be shown that the best hyperplane is found by minimizing $|w|^2 + C(\sum_i \epsilon_i)$, where C is a constant that controls the misclassification cost. It can be shown that [22] this is equivalent to minimizing a dual problem

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j}) \tag{4}$$

subject to

$$\begin{aligned} 0 \leq \alpha_i &\leq C \\ \sum_i \alpha_i y_i &= 0. \end{aligned} \tag{5}$$

If the projected dimensionality is high, then it becomes computationally intensive dealing with terms of the type $\Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$; however, if we have a suitable *kernel* function such that $K(\mathbf{x_i}, \mathbf{x_j}) = \Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$, then we never need to know what Φ is and can solve the optimization problem. Common choices for these kernel functions are polynomial kernels, radial basis functions, and so forth. We note that this linear hyperplane in the high-dimensional space results in a nonlinear separation surface in the original feature space. For more details, see Vapnik [22].

2.4. Learning visual concepts

We now describe the specific approaches for modeling concepts in the different modalities. In case of static visual scenes or objects, the class conditional density functions of the feature vector under the true and null hypotheses are modeled as mixtures of multidimensional Gaussians. Temporal flow is not considered for static objects and scenes. In case of events and objects with spatio-temporal support, we use HMMs with multivariate 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.

In this paper, we compare the performance of GMMs and SVMs for the classification of static scenes and objects. In both cases, the features being modeled are extracted from regions in the video or from the entire frame depending on the type of the concept.

2.5. Learning audio concepts

The scheme for modeling audio-based atomic concepts, such as silence, rocket engine explosion, or music, begins with the annotated audio training set described earlier. Regions corresponding to each class are segmented from the audio and the low-level features extracted. One obvious modeling scheme uses these features to train a GMM for each concept. However, this ignores the duration properties of the audio events; the use of these GMMs to label new (or even training) videos (by assigning each frame in the new data to the most likely generating concept) may yield implausibly short events.

One scheme for incorporating duration modeling is as follows: an HMM is used to model each audio concept and each state in a given HMM has the same observation distribution, namely, the GMM trained in the previous scheme.[2] This can be viewed as imposition of a minimum-duration constraint on the temporal extent of the atomic labels.

Given a set of HMMs, one for each audio concept, during testing (labeling new videos), we use the following schemes to compute the confidences of the different hypotheses.

Scheme 1: We estimate the fractional presence of the different atomic concepts in a shot using the HMMs to generate an N-best list at each audio frame and then average these scores over the duration of the shot.

Scheme 2: We notice that there are variations in the absolute values of these scores due to variations in the shot lengths and the thresholds chosen for generating the N-best list, and so forth. For example, a lower threshold allows for more hypotheses at any one time but also allows a hypothesis to be valid for a longer duration. To counter these variations, we *normalize* these scores by dividing each concept score by the sum of all the concept scores in a particular shot. The scores are now indicative of the relative strengths of the different hypotheses in a given shot rather than their absolute values.

2.6. Representing concepts using speech

Speech cues may be derived from one of two sources: manual transcriptions such as close captioning, where available, or the results of automatic speech recognition (ASR) on the speech segments of the audio. Retrieval of shots relevant to a particular concept is performed in the same manner as standard text-retrieval systems, for example, [24]. Firstly, given

[2] It is closely related to the speech versus nonspeech segmentation scheme of IBM-Spine2, see Kingsbury et al. [23].

transcriptions of either type, the transcriptions must be split into documents and preprocessed ready for retrieval. Documents are defined here in two ways: the words corresponding to a shot or words occurring symmetrically around the center of a shot. (The latter reflects a belief that in highly edited videos, speech cues may occur not just within the unit of the (potentially short) shot, but also in surrounding shots; "surrounding shots" might profitably be defined as "shots in the same scene as the shot of interest," but there is no scene detection in the current system.) This document construction scheme gives a straightforward mapping between documents and shots. The word time marks, necessary to determine the mapping from word tokens to shots, can be obtained using either a forced alignment with an ASR system (for ground truth transcriptions) or directly from the ASR output (for the case of automatically produced transcriptions). The words in each document are then tagged with part of speech (e.g., noun phrase), which enables morphological decomposition to reduce each word to its morph. Finally, stop words are removed using a standard stop-words list.

The procedure for labeling a particular semantic-concept using speech information alone assumes the a priori definition of a set of query terms pertinent to that concept. One straightforward scheme for obtaining such a set of query terms automatically would be to use the most frequent words occurring within shots (or their associated documents) annotated by a particular concept (modulo some stop list, and perhaps incorporating some concept of inverse document frequency); the set might also be derived (or the previous set refined) using human knowledge or word net [25]. Tagging, morphologically analyzing, and applying the stop list to this set of words (in the same way as was applied to the database documents) yield a set of query terms Q for use in retrieving the concept of interest. Retrieval of shots containing the concept then proceeds by ranking documents against Q according to their *Okapi* [7, 24, 26] score, as in standard text retrieval, which provides a ranking of shots.

2.7. Learning multimodal concepts

In the previous sections, we detailed concept models in the individual modalities. Each of these models is used to generate scores for these concepts in unseen video. One or more of these concept scores are then combined or *fused* within models of high-level concepts, which may in turn contribute scores to other high-level concepts. In our current system, this is the step at which information cues from one or more modalities are integrated. (Recall the atomic-concept models used in the system at present that use information from a single modality.)

Assuming a priori definition of the set of intermediate concepts relevant to the higher-level concept, then retrieval of the high-level concepts is a two-class classification problem (concept present or absent) similar to Section 2.4. It is amenable to similar solutions: the modeling of class conditional densities or more discriminative techniques such as SVMs. In this work, the features used in the high-level models will always be scores (as obtained from the atomic-concept models). This is partly to counter the paucity of examples of annotated high-level concepts. In addition, we believe we can build richer models that exploit the interrelationships between atomic concepts, which may not be possible if we model these high-level concepts in terms of their features. We note here that the scores can be likelihood ratios, log likelihoods, SVM classification scores, results of the Okapi formula, and so on. When we fuse the scores in a Bayesian setting, we normalize the scores and effectively treat them as "probabilities" (specifically, posterior probability of a concept, given an observation).

We now detail the two different late-fusion approaches we investigated in this paper. In the first approach, we use a Bayesian network to combine audio, visual, and textual information. Next, we illustrate our novel approach of representing semantic-concepts in terms of a "basis" vector of other semantic-concepts and using a discriminant framework to combine these concept scores together.

2.7.1 Inference using graphical models

A variety of models can be used to model the class conditional distribution of scores; in this work, the models used are Bayesian networks of various topologies and parameterizations. Bayesian networks allows us to graphically specify a particular form of the joint probability density function. Figure 3a represents just one of many possible Bayesian network model structures for integrating scores from atomic-concept models, in which the scores from each intermediate concept are assumed to be conditionally independent given the concept's presence or absence; the parameters for the model (prior to concept presence and the assumed forms of conditional distributions on scores) can be estimated from training data. For example, in Figure 3a, the joint probability function encoded by the Bayesian network, is

$$P(E, A, V, T) = P(E)P(A/E)P(V/E)P(T/E), \tag{6}$$

where E is a binary random variable representing the presence or absence of the high-level concept we are modeling and A, V, and T are the acoustic, visual, and textual scores, respectively.

2.7.2 Classifying concepts using SVMs

In this approach, the scores from all the intermediate concept classifiers are concatenated into a vector, and this is used as the feature in the SVM. This is illustrated in Figure 3b.

We can view the classifiers as nonlinear functions that take points in $\Re^n$ and map them into a scalar, that is, $C(\mathbf{x}) : \Re^n \mapsto \Re$, where $\mathbf{x}$ is an n-dimensional feature vector and C is a classifier that operates on this feature vector. We make the claim that points that are near in the feature space produce similar scores when operated on by these classifiers. This is a reasonable assertion given classifiers such as GMMs, HMMs, and SVMs. Now, if you consider a cluster in the feature space, this maps into a 1-dimensional cluster of scores for any given classifier. If we consider a set of classifiers, the combination of this 1-dimensional cluster of scores will now map into a cluster in this *semantic* feature space. We can then view the SVM for fusion as operating in this new "feature" space and

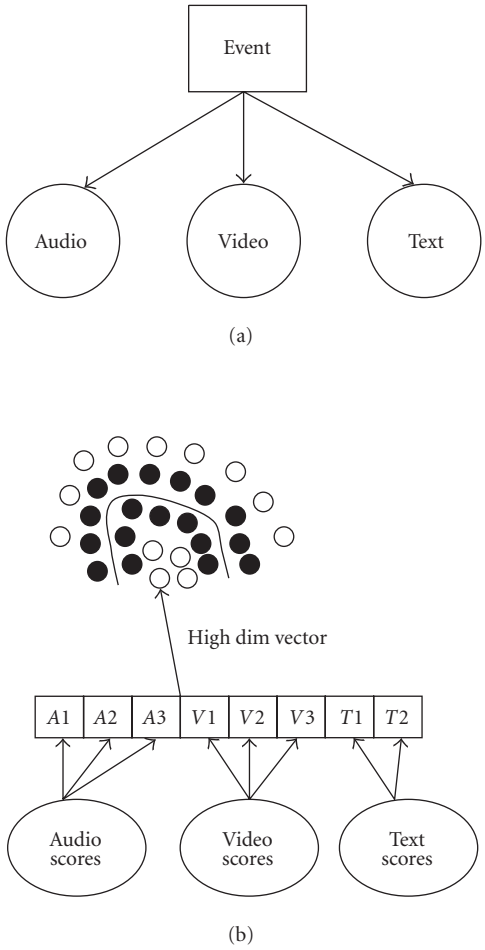


FIGURE 3: Combining information from multiple intermediate concepts (a) Bayesian networks and (b) Support vector machines.

finding a decision boundary. This is illustrated in Figure 4 for a 2-dimensional feature space and 2 classifiers. In this example, there is no advantage in terms of dimensionality reduction going from the 2-dimensional feature space to the 2-classifier "semantic" space. However, in a typical situation, the input feature space can be fairly large compared to the number of classifiers, and here we expect the dimensionality reduction to be useful.

3. EXPERIMENTAL RESULTS

We now demonstrate the application of the semantic-content analysis framework described in Section 2 to the task of detecting several semantic-concepts from the NIST Video TREC 2001 corpus. Annotation is applied at the level of camera shots. We first present results for concepts based on low-level features. These include visual concepts like sky, rocket-object, outdoor, and fire/smoke and audio concepts like speech, music, and rocket engine explosion. We then show how a new concept, rocket launch, could be inferred, based on more than one detected concept.

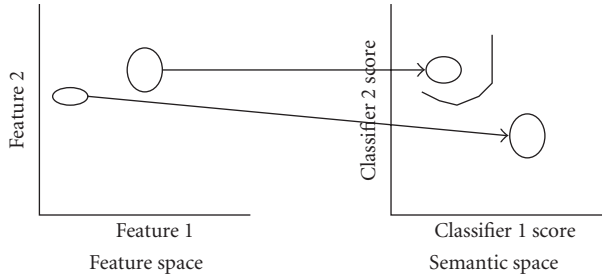


FIGURE 4: Illustration of SVM for fusion.

3.1. The corpus

For the experiments reported in this paper, we use a subset of the NIST Video TREC 2001 corpus, which comprises production videos derived from sources such as NASA and OpenVideo Consortium. Some of the clips contain footage of NASA activities including the space program.

We use 7 videos comprising 1248 video shots. The 7 videos describe NASA activities including its space program. They are sequences entitled anni005, anni006, anni009, anni010, nad28, nad30, and nad55 in the TREC 2001 corpus. The examination of the corpus justifies our hypothesis that the integration of cues from multiple modalities is necessary to achieve good concept labeling or retrieval performance. Of 78 manually annotated rocket launch shots, only 51 contain speech and only a subset of those contain rocket launch related words. The most pertinent audio cues are music and rocket engine explosion, found in 84% and 60% of manually labeled audio samples, respectively. This is due to the highly produced nature of the video content.[3] In the visual side, the rocket shots are from a variety of poses, and in many cases, the rocket exhaust completely occludes the rocket object. Therefore, it seems unlikely that any single audio, speech, or visual cue could retrieve all relevant examples.

3.2. Preprocessing and feature extraction

3.2.1 Visual shot detection and feature extraction

Shot segmentation of these videos was performed using the IBM CueVideo toolkit [27, 28]. Key frames are selected from each shot and low-level features representing color, structure, and shape are extracted.

Color

A normalized, linearized[4] 3-channel HSV histogram is used, with 6 bins each, for hue (H), saturation (S), and 12 bins for intensity (V). The invariance to size, shape, intraframe motion, and their relative insensitivity to noise makes color histograms the most popular features of color content description.

[3] Rocket launch shots that we obtained from TREC video are part of NASA documentaries and typically have such audio and visual overlays.

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

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, a 24-bin histogram of edge directions is obtained as in [29]. The edge direction histogram is a robust representation of shape [30].

Shape

Moment invariants as in Dudani et al. [31] are used to describe the shape of each region. For a binary image mask, the central moments are given by (7),

$$\mu_{pq} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^p (y_i - \overline{y})^q, \tag{7}$$

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

In all, 56 features are extracted to represent the visual properties of the region. Note that regions of interest around objects are specified *manually* during testing and training at present; automating this process for testing is the subject of current research. Note that this may be a simpler problem than object segmentation; we are interested in accurate concept classification, which may be possible without accurate extraction of object contours. A similar set/subset of features can be also obtained at the global level without segmentation and also at difference frames obtained using successive consecutive frames [2].

3.2.2 Audio feature extraction

The low-level features used to represent audio are 24-dim mel-frequency cepstral coefficients (MFCCs), common in ASR systems. MFCCs are typically generated at 10 ms intervals with a sliding window that is 25 ms long. The 25 ms audio sample window is transformed to the frequency domain via Fourier transform and the pitch information is discarded. The Fourier coefficients are then filtered via triangle-shaped band-pass filters (mel filters) that are logarithmically spread in the frequency domain. The resulting filter outputs are then further transformed using a discrete cosine transform (DCT) resulting in MFCCs.

3.3. Lexicon

Our current lexicon comprises more than fifty semantic-concepts for describing events, sites, and objects with cues in audio, video, and/or speech. Only a subset is described in these experiments.

(i) Visual Concepts: rocket object, fire/smoke, sky, outdoor.
(ii) Audio Concepts: rocket engine explosion, music, speech, noise.
(iii) Multimodal Concept: rocket launch.

Of these, the audio concepts and the visual concepts are detected independently and the event *rocket launch* is a high-level concept that is inferred from the detected concepts in multiple modalities.

3.4. Evaluation metrics

The training examples of intermediate audio, visual, and speech concepts have been manually annotated in the corpus. Since data labeled with these events is limited, a cross-validation or leaving-one-sequence-out strategy is adopted in these experiments: models are trained on all-but-one video sequence and tested on the held-out video sequence. The results presented is the combination of this 7-step cross-validation. (Recall that we use a subset of 7 video sequences from the TREC 2001corpus.)

We measure retrieval performance using precision-recall curves. Precision is defined as the number of relevant documents (shots)/total retrieved documents, and recall is defined as the number of relevant documents/the total number of relevant documents in the database. In addition, an overall figure-of-merit (FOM) of retrieval effectiveness is used to summarize performance, defined as average precision over the top 100 retrieved documents.

3.5. Retrieval using models for visual features

We now present results on the detection of the visual concepts using GMMs and SVMs.

3.5.1 Results: GMM versus SVM classification

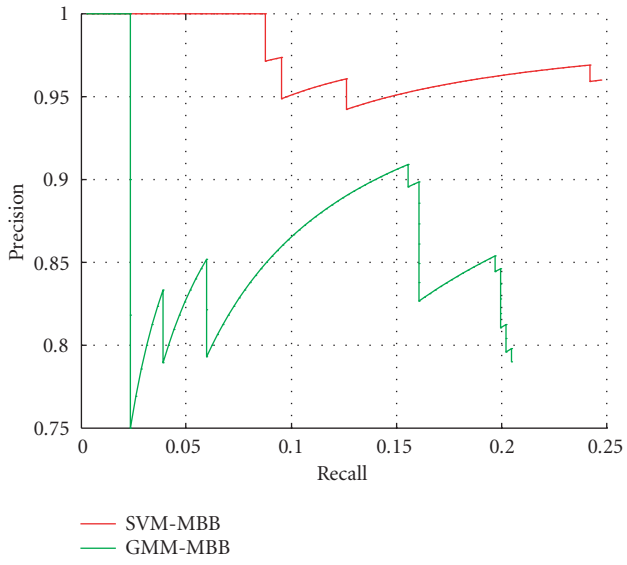
GMM classification builds a GMM for the positive and the negative hypotheses for each feature type (e.g., color histogram, edge direction histogram, etc.) for each semantic-concept. We then merge results across features for these multiple classifiers using the naive Bayes approach. Five Gaussian components are used in each GMM. We note here that we did not experiment with the number of Gaussians in these models. In case of SVM classification, a radial basis function is used with other parameters of the model experimentally chosen.

Table 1 shows the overall retrieval effectiveness for a variety of intermediate (visual) semantic-concepts with SVM and GMM classifiers. Clearly, discriminant classification using SVMs outperforms GMM-based classification. This is because the SVM classifier needs to model less information in terms of what differentiates a positive example from a negative example and therefore requires less data to estimate parameters reliably.

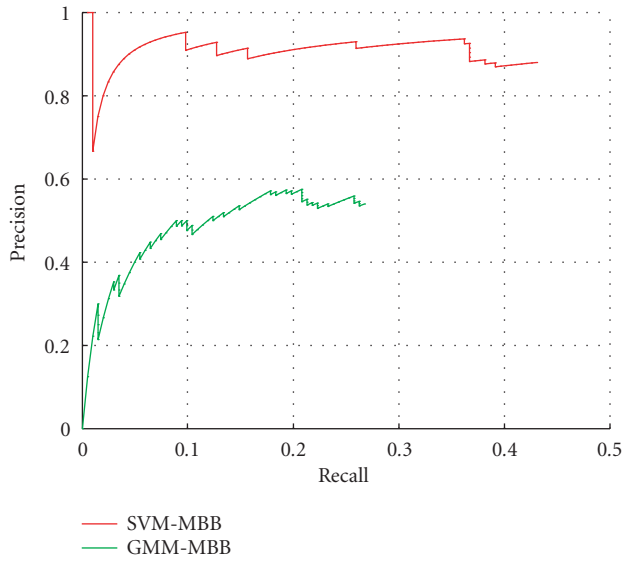
Figures 5 and 6 show the precision-recall curves for 4 different visual concepts using the two classification strategies. Each precision-recall curve compares the performance of the SVM classifier with the GMM classifier for each concept. Since we are interested in the range of recall and precision corresponding to a small number of retrieved items, we limit the plots to depict precision and recall for the first 100 items retrieved. Assuming that 20 items can be seen simultaneously on a screen, this assumes that the user is prepared to scroll through the first five screen shots, which, according to our experience, is a reasonable assumption. We note here that these precision-recall curves (and all other precision-recall

Table 1: Comparing test set accuracy of visual concept classification for the two methods.

Semantic-concept	# Positive examples	SVM : FOM	GMM : FOM
Outdoors	386	0.9727	0.8604
Sky	202	0.9069	0.4454
Rocket	90	0.3854	0.2111
Fire/Smoke	42	0.334	0.1386

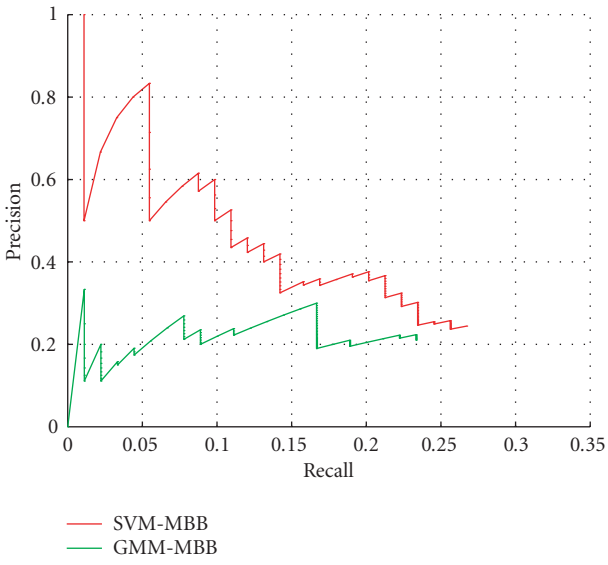


(a)

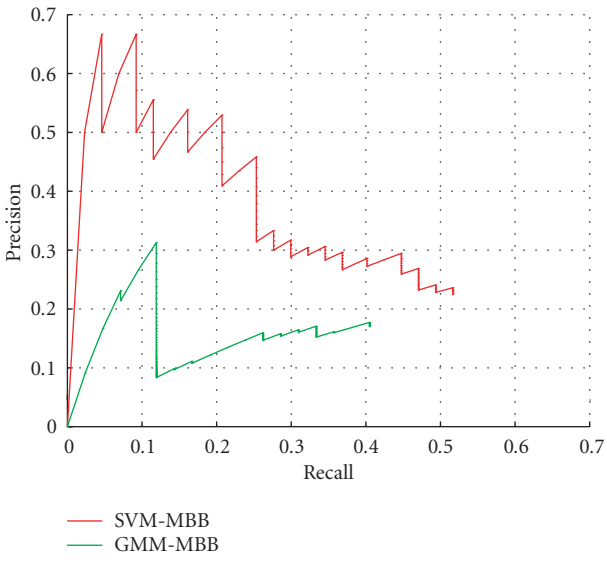


(b)

Figure 5: Precision-recall comparison between SVM and GMM (a) Outdoors (b) Sky.



(a)



(b)

Figure 6: Precision-recall comparison between SVM and GMM (a) Rocket object (b) Fire/smoke.

curves reported in this paper) are interpolated. For example, since we have only 78 ground-truth examples for the rocket launch, we can only have recalls in multiples of 1/78. This implies that for some recall values, we have to estimate the precision in order to get a continuous curve. We interpolate in the document retrieved order. By this we mean the

following: at every retrieved document, we calculate the recall as a fraction of the number of relevant documents retrieved so far and the precision as the fraction of retrieved documents that are relevant. This choice has two implications. Firstly, the density of the curve is low at the low-recall regime and high at the high-recall regime. Secondly, the precision jumps up nonmonotonically every time we get a correct document. This explains the nonmonotonic nature of the graphs. To overcome this, some authors present a noninterpolated precision-recall curve where the precision is calculated only when a relevant document is retrieved and intermediate points are linearly interpolated.

Table 1, Figures 5, 6 bring out a very clear message. As in the case of the concepts *outdoor* and *sky*, when sufficient number of examples is available for training, the SVM classifiers lead to a very accurate retrieval performance with over 90% precision for most of the recall range. Interestingly, even with very small number of examples for training, the SVM classifiers still provide a reasonably accurate detection performance as observed in the case of *rocket object* and *fire/smoke*. In all cases, the SVM classifiers outperform the GMMs.

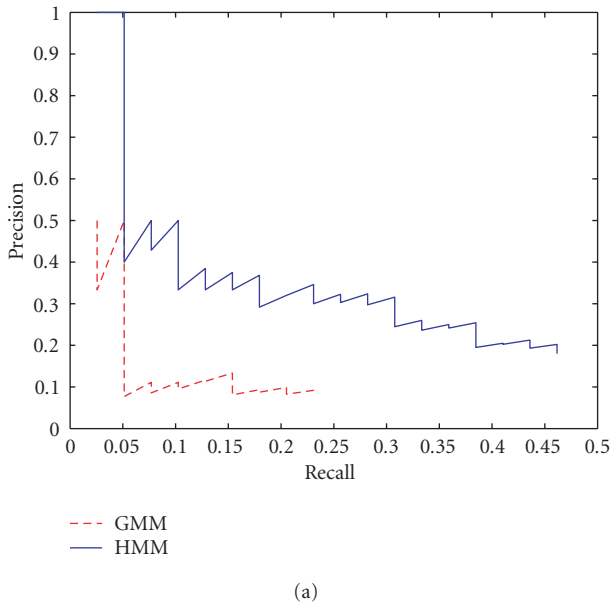
We note here that the rocket object model was highly correlated with rocket launch event. In our experiments, the rocket object model had a better precision-recall performance for rocket launch events compared to rocket object detection. Clearly, this is a case of erroneous annotation. In some shots containing rocket launches, the event was marked but a rocket object was not demarcated, thereby making a shot valid for rocket launch events but not for rocket objects. This is indicative of some of the challenges that we face in relying on an annotated corpus.

3.6. Retrieval using models for audio features

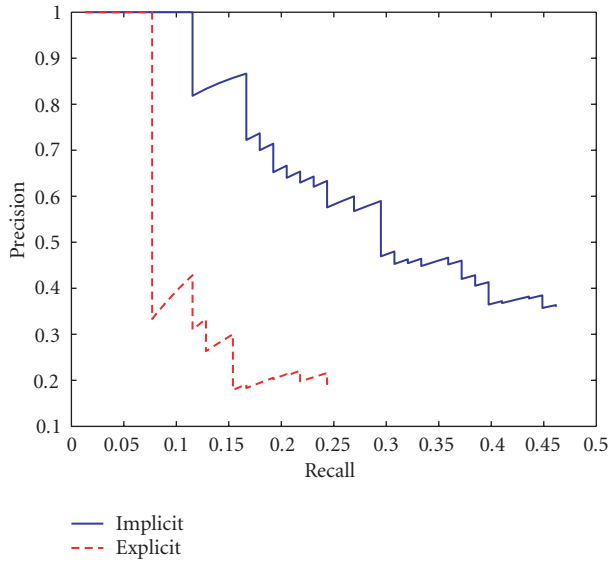
This section presents two sets of results: the first examines the effects of minimum duration modeling upon intermediate concept retrieval and the second examines different schemes for fusing scores from multiple audio-based intermediate concept models in order to retrieve the high-level *rocket launch* concept.

3.6.1 Results: minimum duration modeling

In the first experiment, we study the effect of using a tied-states HMM for duration modeling of a single intermediate concept (rocket engine explosion). The states of the HMM are tied (the output probability distributions at each of the states are identical and is namely the GMM trained on the labeled features for a particular concept). We use a 5-state left-to-right HMM topology. Figure 7a compares the retrieval of the rocket engine explosion concept with HMM and GMM scores, respectively. Notice that the HMM model has significantly higher precision for all recall values compared to the GMM model. Since the minimum duration constraint requires a minimum number of frames to be classified in the same way, this scheme possibly reduces both false positives and false negatives (by not allowing isolated misclassified frames). This effect and the optimal minimum duration length needs to be investigated further. Table 2 shows the FOM for the 4 different audio concepts, using HMMs for the atomic classifiers.



(a)



(b)

Figure 7: (a) Effect of duration modeling. Notice HMM outperforms GMM. (b) Implicit versus explicit fusion. Notice that implicit fusion outperforms explicit fusion.

3.6.2 Results: fusion of scores from multiple audio models

The first approach investigated for score combination is an *implicit* fusion approach (Scheme 2 in Section 2.5), in which the score for a concept is now based on all other concept

Table 2: FOM results: audio retrieval, different intermediate concepts.

Audio model	FOM
Rocket engine explosion	0.38
Music	0.92
Speech	0.89
Speech+Music	0.76

Table 3: FOM results: audio retrieval, GMM versus HMM performance and implicit versus explicit fusion.

Audio model	FOM
GMM (rocket engine explosion)	0.12
HMM (rocket engine explosion)	0.38
Explicit (rocket launch)	0.32
Implicit (rocket launch)	0.56

scores in a given shot

$$F(c_i) = f(c_1, \ldots, c_n) = \frac{\text{Score}\,(c_i)}{\sum_{k=1}^{n} \text{Score}\,(c_k)}. \quad (8)$$

Shot scores for the rocket launch concept are based on the normalized (Scheme 2) score of the rocket engine cue. The second approach investigated is *explicit* fusion, in which we take the scores (from Scheme 1) of rocket engine explosion, music, speech, and speech+music and combine them using a Bayesian network. Figure 7b compares implicit and explicit fusion of the atomic audio concepts for the high-level concept (rocket launch) retrieval. Notice that the implicit fusion scheme has a significantly higher precision for all recall values. This is possibly because of the discriminative nature of the implicit fusion score. It reflects how dominant one audio cue is with respect to the others. Table 3 shows the FOM corresponding to the figures.

3.7. Retrieval using speech

This section presents two sets of results: the retrieval of the rocket launch concept using manually produced ground-truth transcriptions (analogous to closed captioning) and retrieval using transcriptions produced using ASR. For both cases, we investigate the effect of document length upon retrieval performance.

The speech-only retrieval experiments use the subset of video clips from TREC 2001, which have been manually transcribed. For the manually produced transcriptions, words were time-aligned to shots using the IBM HUB4 (broadcast news) ASR system [32]; the speech recognition transcriptions were produced using the same system and the time marks are derived from the ASR output. Prior to generating the ASR, the audio data was preprocessed using an automatic speech/nonspeech segmenter.[5] The frame-level accuracy of the segmenter is 77% (speech 88%, nonspeech 59%). Very short segments are then merged. The ASR error rate over the manually transcribed 13-video retrieval subset is currently 29%. We investigated two schemes for deriving documents from these transcriptions. The first defines a document as comprising words corresponding to a single shot; the second defines documents as the 100 words symmetrically centered on the center of each shot, or the full set of words in the shot if this exceeds 100 words. In either case, there is a one-one correspondence between shots and documents.

Table 4: FOM results: speech retrieval using human knowledge-based query.

Transcription and document type	FOM
Ground truth, shot documents	0.20
Ground truth, 100 word documents	0.15
ASR, shot documents	0.17
ASR, 100 word documents	0.13

Two query term sets Q pertinent to rocket launches were used: the first training set-based query comprises query terms selected from amongst words frequent in rocket launch shots (engines, flight, lift, off, NASA, five, four, three, two, one, shuttle, space) and the second human knowledge-based query is obtained by asking users unfamiliar with the TREC corpus for words expected to be pertinent to the rocket launch event (NASA, ariane, rocket, launch, space, agency, nasda, satellite, spacecraft, space, shuttle, mission). We note that the training set-based query performs better than the latter since it comprises terms chosen with the knowledge of the test set. We chose this query to get a bound on the performance of sophisticated query processing.

Figure 8 and Table 4 show retrieval results when using the human knowledge-based query. Firstly, it appears that shorter documents benefit retrieval in FOM terms and at low recall rates in the graph. The FOM results suggest there may be some benefit from further improving the ASR performance although at very low recall rates, the ASR-shot scheme actually outperforms the ground-truth-shot scheme. We attribute this to poor automatic time-alignments of the manual transcriptions: the videos contain long stretches of music and other nonspeech noise, and in those regions automatic alignment is not as reliable as in speech-only regions. Another important direction for future research appears to be query processing and selection of pertinent query terms: the FOM for ground-truth retrieval using the training set-based query is around 0.30 for both shot- and 100-word-document definitions, whereas Table 4 shows that the comparable FOMs when using the human knowledge-based query is around 0.15.

3.8. Retrieval using fusion of multiple modalities

This section presents results for rocket launch concept which is inferred from concept models based on multiple modalities. As mentioned in Section 2.7, we present results for two different integration schemes.

[5] Using a scheme similar to IBM-Spine2 system and to Scheme 1 of Section 2.5.

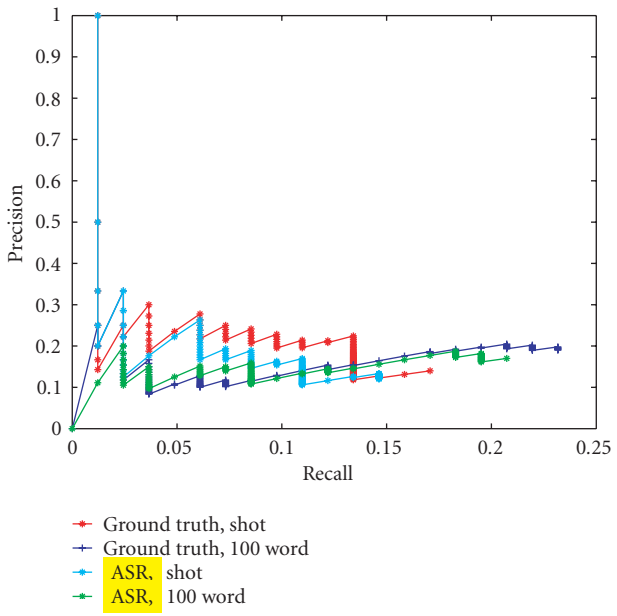


Figure 8: Precision-recall: human knowledge-based query.

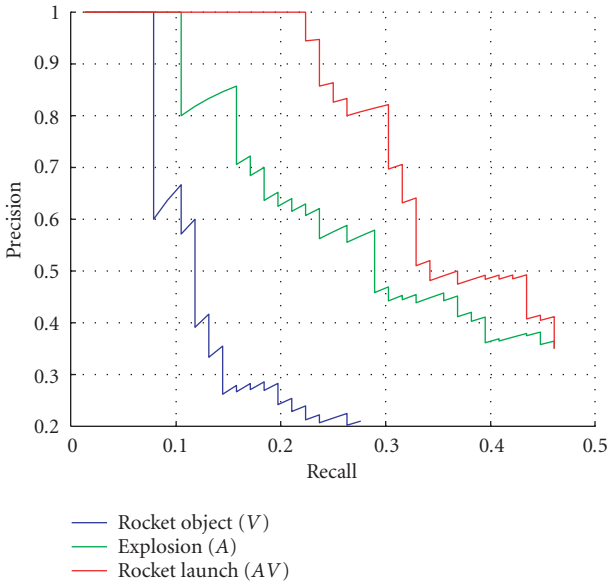


Figure 9: Precision-recall curves for up to 100 retrieved items using the Bayesian network to retrieve video clips depicting rocket launch/take-off.

Table 5: FOM results for unimodal retrieval and the two multimodal fusion models.

Technique	Retrieval FOM
Best unimodal (audio)	0.56
Best visual	0.39
Text (unknown item)	0.14
SVM (audio,text,visual)	0.63
BN (best audio + best visual)	0.62

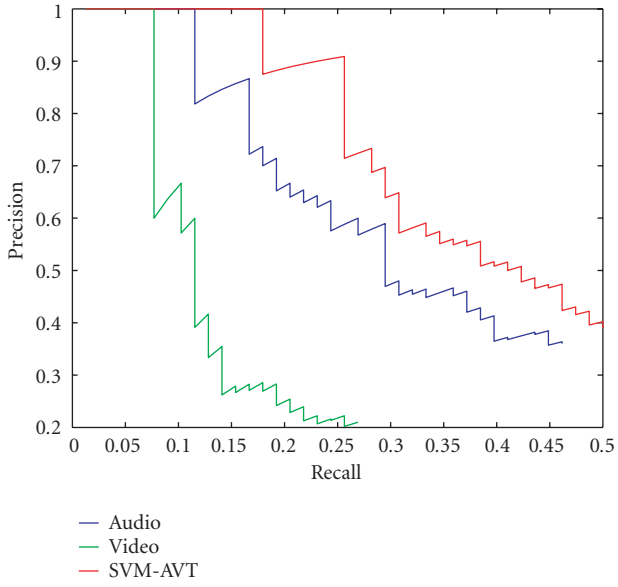


Figure 10: Fusion of audio, text, and visual models using the SVM fusion model for rocket launch retrieval.

3.8.1 Bayesian network integration

A Bayesian network is used to combine the soft decision of the visual classifier for rocket object with the soft decision of the audio classifier for explosion in a model of the rocket launch concept.[6] In this network, all random variables are assumed to be binary valued. During the training phase, we clamp the node E with the ground truth (1 if rocket launch is present in the shot, else 0) while learning the parameters of the network in terms of conditional probability tables. During inference, we present the network with the probability of observing node A and V to be present. The probability of node E taking the value 1 is then inferred. In all cases, the scores emitted by the individual classifiers (rocket object and rocket engine explosion) are processed to fall into the 0–1 range by using the precision-recall curve as a guide. We map acceptable operating points on the precision-recall curve to the 0.5 probability. This is to make maximal and meaningful use of the dynamic range available to us.

Figure 9 illustrates the results of using Bayesian networks for doing fusion. The figure shows the precision-recall performance for the first 100 documents retrieved. Note that the Bayesian network performs much better than either audio or visual models alone.

3.8.2 SVM integration

For fusion using SVMs, we took the scores from all semantic models (Audio: explosion, music, speech, speech-music; Video: rocket, outdoors, sky, fire-smoke; Text: rocket launch), concatenating them into a 9-dimensional feature vector. During SVM training, the class label is clamped to

[6] See Murphy's toolbox [33] for an explanation of soft evidence.

FIGURE 11: The top 20 video shots of rocket launch/take-off retrieved using multimodal detection based on the SVM model. Nineteen of the top 20 are rocket launch shots.

the ground truth (1 for rocket launch shots and −1 for non-rocket launch shots) and the 10-dimensional vector is presented as the observation. The model is cross validated using the aforementioned leave-one-sequence-out approach. Figure 10 presents the precision-recall performance for the SVM fusion and Figure 11 presents qualitative evidence of the success of the SVM fusion approach. In the top 20 retrieved shots, there are 19 rocket launch shots.

Table 5 summarizes the various models in terms of their FOM. Notice that results of both the fusion models are superior to the retrieval results of the individual modalities.

4. CONCLUSIONS

The paper presented an overview of a trainable QBK system for labeling semantic-concepts within unrestricted video. Feasibility of the framework was demonstrated for the semantic-concept rocket launch, first for concept classification using information in single modalities and then for concept classification using information from multiple modalities. These experimental results, whilst preliminary, suffice to show that information from multiple modalities (visual, audio, speech, and potentially video text) can be successfully integrated to improve semantic labeling performance over that achieved by any single modality.

There is considerable potential for improving the schemes described for atomic and high-level concept classification. Future research directions include the utility of multimodal fusion in atomic concept models (using, e.g., coupled HMMs or other dynamic Bayesian networks), and the appropriateness of shot-level rather than scene-level (or other) labeling schemes. Schemes must also be identified for automatically determining the low-level features (from a predefined set of possibilities) which are most appropriate for labeling atomic concepts and for determining atomic concepts (amongst the predefined set of possibilities) which are related to higher-level semantic-concepts. In addition, the scalability of the scheme and its extension to much larger numbers of semantic-concepts must also be investigated.

ACKNOWLEDGMENTS

The authors thank M. Franz, B. Kingsbury, G. Saon, S. Dharanipragada, B. Maison, and other members of the HLT

group at IBM Research. Also B. Tseng and S. Basu of Pervasive Media Management group at IBM Research. M. Slaney of IBM Almaden Research Center for helpful discussions.

REFERENCES

[1] J. R. Smith and S.-F. Chang, "VisualSEEk: a fully automated content-based image query system," in *Proc. 4th ACM International Conference on Multimedia*, pp. 87–98, ACM, Boston, Mass, USA, November 1996.

[2] M. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to Video indexing and retrieval in multimedia systems," in *Proc. IEEE International Conference on Image Processing*, vol. 3, pp. 536–540, IEEE, Chicago, Ill, USA, October 1998.

[3] S. F. Chang, W. Chen, and H. Sundaram, "Semantic visual templates—linking features to semantics," in *Proc. IEEE International Conference on Image Processing*, vol. 3, pp. 531–535, IEEE, Chicago, Ill, USA, October 1998.

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

[5] T. Zhang and C. Kuo, "An integrated approach to multimodal media content analysis," in *Storage and Retrieval for Media Databases 2000*, vol. 3972 of *SPIE Proceedings*, pp. 506–517, SPIE, San Jose, Calif, USA, January 2000.

[6] D. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, Mass, USA, 1996.

[7] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Proc. International Conf. on Computer Vision*, vol. 2, pp. 408–415, IEEE, Vancouver, Canada, July 2001.

[8] M. A. Casey, "Reduced-rank spectra and minimum-entropy priors as consistent and reliable cues for generalized sound recognition," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[9] W. Wolf, "Hidden Markov model parsing of video programs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, pp. 2609–2611, IEEE, Munich, Germany, April 1997.

[10] G. Iyengar and A. B. Lippman, "Models for automatic classification of video sequences," in *Storage and Retrieval for Image and Video Databases VI*, vol. 3312 of *SPIE Proceedings*, pp. 216–227, SPIE, San Jose, Calif, USA, January 1998.

[11] A. M. Ferman and A. M. Tekalp, "Probabilistic analysis and extraction of video content," in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 91–95, IEEE, Kobe, Japan, October 1999.

[12] N. Vasconcelos and A. Lippman, "Bayesian modeling of video editing and structure: semantic features for video summarization and browsing," in *Proc. IEEE International Conference on Image Processing*, vol. 3, pp. 153–157, IEEE, Chicago, Ill, USA, 1998.

[13] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 1331–1334, IEEE, Munich, Germany, April 1997.

[14] B. Adams, C. Dorai, and S. Venkatesh, "Towards automatic extraction of expressive elements from motion pictures: Tempo," in *Proc. IEEE International Conference on Multimedia and Expo*, vol. II, pp. 641–645, IEEE, New York, NY, USA, July 2000.

[15] E. M. Voorhees and D. K. Harman, Eds., *The 10th Text REtrieval Conference (TREC 2001)*, vol. 500-250 of *NIST Special Publication*, NIST, Gaithersburg, Md, USA, 2001.

[16] M. Davis, "Media Streams: an iconic visual language for video annotation," *Telektronikk*, vol. 89, no. 4, pp. 59–71, 1993.

[17] C. Neti, G. Potamianos, J. Luettin, et al., "Audio-visual speech recognition," Final workshop 2000 report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, Md, USA, October 2000.

[18] G. Iyengar and C. Neti, "Speaker change detection using joint audio-visual statistics," in *Proc. RIAO*, Paris, France, April 2000.

[19] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[20] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1st edition, 1993.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[22] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA, 1995.

[23] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The IBM SPINE-2 evaluation system," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Orlando, Fla, USA, May 2002.

[24] M. Franz and S. Roukos, "TREC-6 Ad-hoc retrieval," in *Proc. 6th Text REtrieval Conference (TREC-6)*, vol. 500-240 of *NIST Special Publication*, pp. 511–516, NIST, Gaithersburg, Md, USA, 1998.

[25] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, M. I. T. Press, Cambridge, Mass, USA, 1998.

[26] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *The 3rd Text REtrieval Conference (TREC-3)*, vol. 500-225 of *NIST Special Publication*, pp. 109–126, NIST, Gaithersburg, Md, USA, 1995.

[27] IBM Almaden Research Center, "The IBM cuevideo project," 1997, in http://www.almaden.ibm.com/projects/cuevideo.shtml.

[28] J. R. Smith, S. Srinivasan, A. Amir, et al., "Integrating features, models, and semantics for TREC video retrieval," in *Proc. 10th Text REtrieval Conference (TREC 2001)*, vol. 500-250 of *NIST Special Publication*, pp. 240–249, NIST, Gaithersburg, Md, USA, 2001.

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

[30] A. K. Jain, A. Vailaya, and W. Xiong, "Query by video clip," *Multimedia Systems: Special Issue on Video Libraries*, vol. 7, no. 5, pp. 369–384, 1999.

[31] S. Dudani, K. Breeding, and R. McGhee, "Aircraft identification by moment invariants," *IEEE Trans. on Computers*, vol. 26, no. 1, pp. 39–45, 1977.

[32] R. Bakis, S. Schen, P. Gopalakrishnan, R. Gopinath, S. Maes, and L. Polymenakos, "Transcription of broadcast news—system robustness issues and adaptation techniques," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 711–714, IEEE, Munich, Germany, April 1997.

[33] K. Murphy, "Bayes net toolbox for matlab," 2001, http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html.

W. H. Adams joined the T. J. Watson Research Center as a Senior Programmer in 1990. He received his Master's degree in computer science from the University of Illinois at Champaign in 1973. While at IBM, he contributed to a variety of projects, including the Power Visualization System, Air Traffic Control Using Speech Game, and Literacy Tutor Using Speech. Prior to joining IBM, Mr. Adams was Manager of the Digital Equipment Corporation VMS workstation group and project leader of X-11 device-independent server. He has an extensive background in video games as a developer and manager. He was director of game development at Bally Midway. Among the video games he developed is TRON, which was awarded Coin-Operated Game of the Year by the Electronic Games Magazine. Mr. Adams has worked as a contractor, creating a number of embedded system products, and he developed applications for police and fire departments at the Motorola Communications Division.

Giridharan Iyengar has been a research staff member in the Audiovisual Speech Technologies Group at the IBM T. J. Watson Research Center since 1999. He received his B.Tech in electrical engineering from the Indian Institute of Technology, Mumbai in 1990. After working as an Engineer in Larsen and Toubro, India, for one year, he obtained his Master's degree in electrical engineering from the University of Ottawa, Canada. He then did his doctoral work at the MIT Media Laboratory, where he worked on video retrieval and indexing. Mr Iyengar is a member of the IBM team that participates in TREC video track organized by NIST. In the past year, he has been the project leader of the Multimedia Mining Project at IBM Research. He has authored over 30 papers, filed 8 patents (4 currently issued), participated in program committees of conferences, and reviewed for journals in multimedia, image processing, and computer vision and image understanding. His research interests include multimodal signal processing, video indexing and retrieval, speech processing, and information retrieval.

Ching-Yung Lin received the B.S. and M.S. degrees from National Taiwan University in 1991 and 1993, respectively, and his Ph.D. degree from Columbia University in 2000, all in electrical engineering. He was a Wireless 2nd Lieutenant at Taiwan Air Force, Taiwan, 1993–1995, and an instructor of Network Communication Lab at National Taiwan University, 1995–1996. Since 2000, he has been a research staff member in the IBM T. J. Watson Research Center, New York. His current research interests include multimedia semantic analysis and understanding, multimedia security, and multimedia transmission and networking. Dr. Lin was the primary contributor in the design of the first successful multimedia-content authentication system and in the design of the first public watermarking system surviving pring-and-scan process. He has been serving as panelist, technical committee member, and invited speaker at various IEEE/ACM/SPIE conferences. He was the special session organizer of multimedia security in IEEE ITCC 2001, and the publicity chair of the IEEE PCM 2001. He is the Technical Program Cochair of the IEEE ITRE 2003 and will serve as a guest coeditor of the proceedings of the IEEE special issue on Digital Rights Management in 2004. Dr. Lin is the recipient of Lung-Teng Thesis Award and an Outstanding Paper Award in CVGIP. Dr. Lin is the author/coauthor of 40+ journal and conference papers. He holds three US patents and five pending patents in the fields of multimedia security and multimedia semantic analysis.

Milind Ramesh Naphade received his B.E. degree in instrumentation and control engineering from the University of Pune, India, in July 1995, ranking first among the university students in his discipline. He received his M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign 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 from August 1996 to March 2001. In 2001, he joined the Pervasive Media Management Group at the IBM T. J. Watson Research Center in Hawthorne, NY, as a research staff member. He has worked with the Center for Development of Advanced Computing (C-DAC) in Pune, India, from July 1994 to July 1996, in the Applications Development Group. He has worked with the Kodak Research Laboratories of the Eastman Kodak Company in the summer of 1997 and with the Microcomputer Research Laboratories at Intel Corporation in the summer of 1998. He is a member of the IEEE and the honor society of Phi Kappa Phi. He has published over 35 research articles, publications, and book chapters in the field of media analysis and learning. His research interests include audiovisual 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.

Chalapathy Neti received his B.S. degree in electrical engineering from the India Institute of Technology, Kanpur, India, his M.S. degree in electrical engineering from Washington State University, and his Ph.D. degree in biomedical engineering from The Johns Hopkins University. He is currently a Research Manager in the Human Language Technologies Department at the IBM T. J. Watson Research Center in Yorktown Heights, New York. In this position, he is leading a group of researchers in developing algorithms and technologies for the joint use of audio and visual information for robust speech recognition, speaker recognition, and multimedia content analysis and mining. He has been with IBM since 1990. Dr. Neti's main research interests are in the area of perceptual computing (using a variety of sensory information sources to recognize humans, their activity, and intent), speech recognition, multimodal conversational systems for information interaction, and multimedia content analysis for search and retrieval. He has authored over 30 articles in these fields and has five patents and several pending. He is an active member of the IEEE, a member of the IEEE Multimedia Signal Processing technical committee, and an Associate Editor of the IEEE transactions on Multimedia.

Harriet J. Nock is a Researcher in the Audiovisual Speech Technologies Group at the IBM T. J. Watson Research Center. She is a team member of the Multimedia Mining Research Project at IBM Research and a member of the IBM team that participates in TREC video track. She holds a B.S. in computer science (1994), an MPhil in computer speech and language processing (1996), and a Ph.D. in information engineering (2001), all from the University of Cambridge, UK. She has taken part in the Johns Hopkins University Summer Research Workshop on Large-Vocabulary Speech Recognition (1997) and has held Visiting Research Fellow positions at the Center for Language and Speech Processing at the Johns Hopkins University (1998) and at the Signal, Speech and Language Interpretation Lab at the University of Washington (2000–2001). Her research interests currently include automatic speech recognition, multimodal signal processing, and multimodal information retrieval.

John R. Smith works for IBM Research at the Watson Laboratories in New York where he manages the Pervasive Media Management Group.

23

COMPOSITION COMMENTS

1 We changed "Gaussian Mixtures" to "Gaussian mixtures models". Please

2 Please note that the number of keywords should not exceed six words.

3 We added the highlighted delimiter. Please check.

4 We changed "Gaussian Mixture Models, Hidden Markov Models" to "GMMs, HMMs", respectively, throughout except in their first occurrences. Please check.

5 We added "SVMs" after "support vector machines" in its first occurrence, while we will use the acronym "SVMs" throughout.

6 We changed "richer models can be built" to "we can build richer models". Please check.

7 Should we change "the figures" to "Figures 7a and 7b"? Please check.

8 We added "image" to the title and changed the last page number in [1]. Please check.

9 We added "video" to the title in [2]. Please check.

10 We changed the format of the booktitle in [2]. Please check.

11 We changed the volume number and the range of pages in [12]. Please check.

12 We updated the information of [15]. Please check.

13 we changed the format of the journal name in [16]. Please check.

14 We changed the publisher name and the year of publication in [19]. Please check.

15 We changed the volume number in [26]. Please check.

16 We changed the URL of [27]. Please check.

17 We added the range of pages in [28]. Please check.

18 We added "for matlab" instead of "(bnt)" in [33]. Please check.

19 The permenent URL in [33] has been changed to the one we added. Please check.

20 We changed "Bill Adams" to "W. H. Adams". Please check.

21 Since you have differentiated between the curves in Figures

5, 6, 7, 9, and 10 by colors, they should be printed in colors. Otherwise, please use another way to differentiate between them (e.g., solid line, dashed line, dotted-dashed line...).

22 Please note that the colors are essential in Figure 8, accordingly, it should be printed in colors.

23 Please provide more detailed biography for "John R. Smith" that should not exceed 200 words.