Generation and dissemination of digital media content poses a challenging problem of efficient storage and retrieval. Of particular interest to us are audio and visual content. From sharing of picture albums and home videos to movie advertisement through interactive preview clips, live broadcasts of various shows or multimedia reports of news as it happens, multimedia information has found in the internet and the television powerful media to reach us. With innovations in hand-held and portable computing devices and wired and wireless communication technology (pocket PCs, organizers, cell-phones) on one end and broadband internet devices on the other, supply and dissemination of unclassified multimedia is overwhelming. Humans assimilate content at a semantic level and apply their knowledge to the task of sifting through large volumes of multimodal data. To invent tools that can gain widespread popularity we must try to emulate human assimilation of this content. We are thus faced with the problem of multimedia understanding if we are to bridge the gap between media features and semantics.

Current techniques in content-based retrieval for image sequences support the paradigm of query by example using similarity in low-level media features \[1,2,3,4,5,6\]. The query must be phrased in terms of a video clip or at least a few key frames extracted from the query clip. Retrieval is based on a matching algorithm, which ranks the database clips according to a heuristic measure of similarity between the query and the target. While effective for browsing and low-level search, this paradigm has limitations. Low-level similarity may not match with the user’s perception of similarity. Also, the assumption that clips
reflecting desire are available during query is unrealistic. It is also essential to fuse information from multiple modalities, especially the image sequence and audio streams. Most systems use either the image sequence \cite{5,6,4,2,1}, or the audio track \cite{7,8,9,10,11,12}, while few use both the modalities \cite{13,14,12}.

One way of organizing a video for efficient browsing and searching is shown in Figure 1. A systematic top-down breakdown of the video into scenes, shots and key frames exists in the form of a table of contents (ToC). To enable access to the video in terms of semantic concepts, there needs to be a semantic index (SI). The links connect entries in the SI to shots/scenes in the ToC and also indicate a measure of confidence.

![Figure 1: Organizing a Video with a Table of Contents (ToC) and a Semantic Index (SI). The ToC gives a top-down break-up in terms of scenes, shots and key frames. The SI lists key-concepts occurring in the video. The links indicate the exact location of these concepts and the confidence measure.](image)

Automatic techniques for generating the ToC exist, though they use low-level features for extracting key frames as well as constructing scenes. The first step in generating the ToC is the segmentation of the video track into smaller units. Shot boundary detection can be performed in compressed domain \cite{15,16,17} as well as uncompressed domain \cite{18}. Shots can be grouped based on continuity, temporal proximity and similarity to form scenes \cite{5}. Most systems support query by image sequence content \cite{2,3,4,5,6} and can be used to group shots and enhance the ability to browse. Naphade et al. \cite{14} presented a scheme, that supports query by audiovisual content using dynamic programming. The user may browse a video and then provide one of the clips in the ToC structure as an example to drive the retrieval systems mentioned earlier. Chang et al. \cite{2} allow the user to provide a sketch of a dominant object along with its color shape and motion trajectory. Key frames can be extracted from shots to help efficient browsing.

The need for a semantic index is felt to facilitate search using key words or key concepts. To support such semantics, models of semantic concepts in terms of multimodal representations are needed. For example, a query to find explosion on a beach can be supported if models for the concepts explosion and beach are
represented in the system. This is a difficult problem. The difficulty lies in the
gap, that exists between low-level media features and high-level semantics.
Querying using semantic concepts has motivated recent research in semantic video
indexing [13,19,20,12] and structuring [21,22,23]. We [13] presented novel ideas
in semantic indexing by learning probabilistic multimedia representations of
semantic events like explosion and sites like waterfall [13]. Chang et al. [19]
introduced the notion of semantic visual templates. Wolf et al. [21] used hidden
Markov models to parse video. Ferman et al. [22] attempted to model semantic
structures like dialogues in video.

The two aspects of mapping low-level features to high-level semantics are the
concepts represented by the multiple media and the context, in which they
appear. We view the problem of semantic video indexing as a multimedia
understanding problem. Semantic concepts do not occur in isolation. There is
always a context to the co-occurrence of semantic concepts in a video scene. We
presented a probabilistic graphical network to model this context [24,25] and
demonstrated that modeling the context explicitly, provides a significant
improvement in performance. For further details on modeling context, the reader
is referred to [24,25]. In this paper we concentrate on the problem of detecting
complex audiovisual events. We apply a novel learning architecture and
algorithm to fuse information from multiple loosely coupled modalities to detect
audiovisual events such as explosion.

Detecting semantic events from audio-visual data with spatio-temporal support
is a challenging multimedia understanding problem. The difficulty lies in the gap
that exists between low-level features and high-level semantic labels. Often, one
needs to depend on multiple modalities to interpret the semantics reliably. This
necessitates efficient schemes, which can capture the characteristics of high
level semantic events by fusing the information extracted from multiple
modalities.

Research in fusing multiple modalities for detection and recognition has
attracted considerable attention. Most techniques for fusing features from
multiple modalities, having temporal support are based on Markov models.
Examples include the hidden Markov model (HMM) [26] and several variants of
the HMM, like the coupled hidden Markov model [27], factorial hidden Markov
model [28], the hierarchical hidden Markov model [13] etc. A characteristic of
these models is the stage, at which the features from the different modalities are
merged.

We present a novel algorithm, which combines feature with temporal support
from multiple modalities. Two main features that distinguish our model from
existing schemes are (a) the ability to account for non-exponential duration and
(b) the ability to map discreet state input sequences to decision sequences. The
standard algorithms modeling the video-events use HMMs which models the
duration of events as an exponentially decaying distribution. However, we argue
that the duration is an important characteristic of each event and we
demonstrate it by the improved performance over standard HMMs. We test the
model on the audio-visual event explosion. Using a set of hand-labeled video
data, we compare the performance of our model with and without the explicit
model for duration. We also compare performance of the proposed model with
the traditional HMM and observe that the detection performance can be improved.

2. PRIOR ART

In this section we review prior art in the fields of event detection and information fusion using multimedia features.

2.1 EVENT DETECTION

Recent work in temporal modeling of image sequences includes work in parsing and structuring as well as modeling visual events. Statistical models like the hidden Markov models (HMM) have been used for structuring image sequences [5,21,22]. Yeung et al. introduced dialog detection [5]. Topical classification of image sequences can provide information about the genres of videos like news, sports, etc. Examples include [29]. Extraction of semantics from image-sequences is difficult. Recent work dealing with semantic analysis of image sequences include Naphade et al. [13,30], Chang et al. [19], and Brand et al. [27]. Naphade et al. [13,30] use hidden Markov models to detect events in image sequences. Chang et al. [2] allow user-defined templates of semantics in image sequences. Brand et al. [27] use coupled HMMs to model complex actions in Tai Chi movies.

Recent work in segmentation and classification of audio streams includes [7,8,9,10,11,12,31,32]. Naphade and Huang [7] used hidden Markov models (HMMs) for representing the probability density functions of auditory features computed over a time series. Zhang and Kuo [12] used features based on heuristics for audio classification. HMMs have been successfully applied in speech recognition.

Among the state of the art techniques in multimedia retrieval very few techniques use multiple modalities. Most techniques, using audiovisual data perform temporal segmentation on one medium and then analyze the other medium. For example the image sequence is used for temporal segmentation and the audio is then analyzed for classification. Examples include [33,34], and the Informedia project [35] 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 [36], sports [12,29,37], meeting videos etc. Wang et al. [34] survey a few techniques for analysis using a similar approach for similar domains. In case of domain-independent retrieval, while existing techniques attempt to determine what is going on in the speech-audio, most techniques go as far as classifying the genre of the video using audiovisual features. Other techniques for video analysis include the unsupervised clustering of videos [38]. Naphade et al. [14] have presented an algorithm to support query by audiovisual content. Another popular domain is the detection and verification of a speaker using speech and an image sequence obtained by a camera looking at the person [39]. This is particularly applicable to the domain of intelligent collaboration and human-computer interaction. Recent work in semantic video indexing includes Naphade et al. [13,40,41].
2.2 FUSION MODELS

Audio-Visual analysis to detect the semantic concepts in videos pose a challenging problem. One main difficulty arises from the fact that the different sensors are noisy in terms of the information they contain about different semantic concepts. For example, based on pure vision, it’s hard to make out between an explosion and normal fire. Similarly, audio alone may give confusing information. On one hand one may be able to filter out the ambiguity arising from one source of information by looking (analysing) another source. While, on the other hand, these different source may provide complimentary information which may be essential in inference.

Motivated by these difficulties, in the past few years a lot of research has gone into developing algorithms for fusing information from different modalities. Since different modalities may not be sampled at the same temporal rate, it becomes a challenging problem to seamlessly integrate different modalities (e.g. audio is normally sampled at 44KHz whereas video is sampled at 30fps.) At the same time, one may not even have the synchronized streams (sources of information) or the sources of information may have very different characteristics (audio - continuous, inputs to the computer through keyboard - discrete.) If we assume that one can get features from the different streams on a common scale of time, the two main categories of fusion models are those that favor early integration of features versus those that favor late integration. Early integration refers to combining the information at the level of raw features. Simple early integration is often observed in the form of concatenation of weighted features from different streams. More involved models of early integration have been proposed by using some form of Markov models. [27] have proposed the coupled hidden Markov models and used it for detection of human activities. Ghahramani et. al. [28] have proposed the factorial hidden Markov models. The main difference in these models arises from the conditional independence assumptions that they make between the states of the different information sources. They assume that the different sources are tightly coupled and model them using a single generative process.

In many situations, especially when the different sources are providing complimentary information one may prefer late integration. It refers to doing inferencing of each stream independently of the others and then combining the output of the two. This is especially important and shows improved results as now one essentially looks at the essential information contained in the different streams and the sensor depend characteristics do not play any role. It also allows one to learn different models for each source independently of one another and then combine the output. One may simply look at the weighted decisions of different sources or may actually use probabilistic models to model the dependencies. For example [42] have proposed the use of dynamic Bayesian networks over the output of the different streams to solve the problem of speaker detection. Similarly, [13] have proposed the use of hierarchical HMMS.

We observed that in the case of movie, the audio and the visual streams normally carry complimentary information. For example, a scene of explosion is not just characterized by a huge thunder but also a visual effect corresponding to bright red and yellow colors. Motivated by this fact we propose the use of late coupling which seems to suit better for this framework. Fusion of multimodal feature streams (especially audio and visual feature streams) has been applied to
problems like Bimodal speech [43], speaker detection [42], summarization of video [36], query by audio-visual content [14] and event detection in movies [13]. Examples of fusion of other streams include fusion of text and image content, motion and image content etc.

3. PROBABILISTIC MODELING OF MEDIA FEATURES

We presented a probabilistic architectures of multiject models representing sites, objects and events for capturing semantic representations [13,24]. Bayes decision theory [44] and statistical learning form the core of our architecture. We briefly review the characteristics and the assumptions of this architecture.

3.1 PROBABILISTIC MULTIMEDIA OBJECTS (MULTIJETS)

Semantic concepts (in video) can be informally categorized into objects, sites and events. Any part of the video can be explained as an object or an event occurring at a site or location. Such an informal categorization is also helpful for selecting structures of the models, which are used to represent the concepts. For the automatic detection of such semantic concepts we propose probabilistic multimedia objects or multijets.

A multiject represents a semantic concept that is supported by multiple media features at various levels (low level, intermediate level, high level) through a structure that is probabilistic [13,30]. Multijets belong to one of the three categories: objects (car, man, helicopter), sites (outdoor, beach), or events (explosion, man-walking, ball-game). Figure 2 illustrates the concepts of a multiject.

\[
P(\text{Outdoor}=\text{Present}|\text{Multimedia features, other multijets})=0.7 \\
P(\text{Outdoor}=\text{Absent}|\text{Multimedia features, other multijets})=0.3
\]

Figure 2: A probabilistic multimedia object (multiject).

A multiject is a flexible, open-ended semantic representation. It draws its support from low-level features of multiple media including audio, image, text, and closed caption [13]. It can also be supported by intermediate-level features, including semantic templates [2]. It can also use specially developed high-level feature detectors like face detectors. A multiject can be developed for a semantic concept if there is some correlation between low-level multimedia features and high-level semantics. In the absence of such correlation, we may not be able to learn a sufficiently invariant representation. Fortunately many semantic
concepts are correlated to some multimedia features, and so the framework has the potential to scale.

Multijects represent semantic concepts, that have static as well as temporal support. Examples include sites like sky, water-body, snow, outdoor, explosion, flying helicopter, etc. Multijects can exist locally (with support from regions or blobs) or globally (with support from the entire video frame). The feature representation also corresponds to the extent of the spatiotemporal support of the multiject. We have described techniques for modeling multijects with static support in Naphade and Huang [24]. In this paper we concentrate on multiject models for events with temporal support in media streams.

3.2 ASSUMPTIONS

We assume that features from audiovisual data have been computed and refer to them as $X$. We assume that the statistical properties of these features can be characteristic signatures of the multijects. For distinct instances of all multijects, we further assume, that these features are independent identically distributed random variables drawn from known probability distributions, with unknown deterministic parameters. For the purpose of classification, we assume that the unknown parameters are distinct under different hypotheses and can be estimated. In particular, each semantic concept is represented by a binary random variable. The two hypotheses associated with each such variable are denoted by $H_i$, $i \in \{0,1\}$, where 0 denotes absence and 1 denotes presence of the concept. Under each hypothesis, we assume that the features are generated by the conditional probability density function $P_i(X)$, $i \in \{0,1\}$. In case of site multijects, the feature patterns are static and represent a single frame. In case of events, with spatiotemporal support, $X$ represents a time series of features over segments of the audiovisual data. We use the one-zero loss function [45] to penalize incorrect detection. This is shown in Equation 1:

$$
\lambda(\alpha_i | w_i) = 0 \quad i = j
$$

$$
= 1 \quad i \neq j
$$

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

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

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

(2)
The term \( p(x|\omega_j) \) is the likelihood of \( \omega_j \) and the test based on the ratio in Equation (2) is called the likelihood ratio test \([44, 45]\).

### 3.3 Multiject Models for Audio and Visual Events

Interesting semantic events in video include explosion, car chase etc. Interesting semantic events in audio include speech, music, explosion, gunshots, etc. We propose the use of hidden Markov models for modeling the probability density functions of media features under the positive and negative hypotheses.

We model a temporal event using a set of states with a Markovian state transition and a Gaussian mixture observation density in each state. We use continuous density models in which each observation probability distribution is represented by a mixture density. For state \( j \) the probability \( b_j(o_t) \) of generating observation \( o_t \) is given by Equation (3):

\[
b_j(o_t) = \sum_{m=1}^{M_j} c_{jm} N(o_t; \mu_{jm}, \Sigma_{jm})
\]

where \( M_j \) is the number of mixture components in state \( j \), \( c_{jm} \) is the weight of the \( m^{th} \) component, and \( N(o_t; \mu, \Sigma) \) is the multivariate Gaussian with mean \( \mu \) and covariance matrix \( \Sigma \) as in Equation (4):

\[
N(o_t; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o_t-\mu)\Sigma^{-1}(o_t-\mu)}
\]

The parameters of the model to be learned are the transition matrix \( A \), the mixing proportions \( c \), and the observation densities \( b \). With \( q_t \) denoting the state at instant \( t \) and \( q_{t+1} \) the state at \( t+1 \), elements of matrix \( A \) are given by \( a_{ij} = \Pr(q_{t+1} = j | q_t = i) \). The Baum-Welch re-estimation procedure \([46, 26]\) is used to train the model and estimate the set of parameters. Once the parameters are estimated using the training data, the trained models can then be used for classification as well as state sequence decoding \([46, 26]\). For each event multiject, a prototype HMM with three states and a mixture of Gaussian components in each state is used to model the temporal characteristics and the emitting densities of the class. For each mixture, a diagonal covariance matrix is assumed.

We have developed multijects for audio events like human-speech, music \([7]\) and flying helicopter \([40]\), and visual events like explosion \([13]\) etc.

### 4. A Hierarchical Fusion Model

In this paper, we discuss the problem of fusing multiple feature streams enjoying spatio-temporal support in different modalities. We present a hierarchical fusion model (see Figure 8), which makes use of late integration of intermediate decisions.
To solve the problem we propose - Hierarchical duration dependent input output Markov model. There are four main considerations that have led us to the particular choice of the fusion architecture.

- We argue that, the different streams contain information which is correlated to one another only at high level. This assumption allows us to process output of each source independently of one another. Since these sources may contain information which has highly temporal structure, we propose the use of hidden Markov models. These models are learned from the data and then we decode the hidden state sequence, characterizing the information describing the source at any given time. This assumption leads us to a hierarchical approach.

![Figure 3: Consider the multimodal concept represented by node E and multimodal features represented by nodes A and V in this Bayesian network. The network implies that features A and V are independent given concept E.](image)

- We argue that the different sources contain independent information. At high level, the output of one source is essentially independent of the information contained in the other. However, conditioned upon a particular bi-modal concept, these different sources may be dependent on one another. This suggests the use of an alternative dependence assumption contrary to conventional causal models like HMMs. This is illustrated in Figures 3 and 4.

![Figure 4: Consider the multimodal concept represented by node E and multimodal features represented by nodes A and V in this Bayesian network. The network implies that features A and V are dependent given concept E.](image)

Note the difference between the assumptions implied by Figures 3 and 4. This idea can be explained with an example event, say an explosion. Suppose the node E represents event explosion and the nodes A and V represent some high characterization of audio and visual features. According to figure 3 given the fact that an explosion is taking place, the
audio and visual representations are independent. This means that once we know an explosion event is taking place, there is no information in either channel for the other. On the contrary Figure 4 implies that given an explosion event the audiovisual representations are dependent. This means that the audio channel and video channel convey information about each other only if the presence of event explosion is known. These are two very different views to the problem of modeling dependence between multiple modalities.

**Figure 5:** This figure illustrates the Markovian transition of the input output Markov model. Random variable $E$ can be present in one of the two states. The model characterizes the transition probabilities and the dependence of these on input sequence $y$.

- We want to model temporal events. We therefore need to expand the model in Figure 4 to handle temporal dependency in the event $E$. We make the usual Markovian assumption thus making the variable $E$ at any time dependent only on its value at the previous time instant. This leads to an input output Markov model (IOMM). This is shown in Figure 5.

- Finally, an important characteristic of semantic concepts in videos is their duration. This points to the important limitation of hidden Markov models. In HMMs the probability of staying in any particular state decays exponentially. This is the direct outcome of the one Markov property of these models. To alleviate this problem, we explicitly model these probabilities. This leads us to what we call duration dependent input output Markov model (DDIOMM) as shown in Figure 6.
**Figure 6:** This figure illustrates the Markovian transition of the input output Markov model along with duration models for stay in each state. Random variable $E$ can be present in one of the two states. The model characterizes the transition probabilities and the dependence of these on input sequence $y$ and duration of stay in any state. Note that missing self transition arrows.

Figure 7, shows the temporally rolled out Hierarchical model proposed based. This model takes as input, data from multiple sensors, which in the case of videos reduces to two streams Audio and Video. Figure 8 shows these two streams with one referring to audio features or observations $A_{O1}, ..., A_{OT}$ and the other to video features $V_{O1}, ..., V_{OT}$. The media information represented by these features is modeled using hidden Markov models which here are referred to as media HMM as these are used to model each media. Each state in the media HMMs represents a stationary distribution and by using the Viterbi decoder over each feature stream, we essentially cluster features spatio-temporally and quantize them through state identities. State-sequence-based processing using trained HMMs can be thought of as a form of guided spatio-temporal vector quantization for reducing the dimensionality of the feature vectors from multiple streams [13].

Once the individual streams have been modeled, inferencing is done to decode the state sequence for each stream of features. More formally, consider $S$ streams of features. For each stream, consider feature vectors $f_{1}^{(s)}, ..., f_{T}^{(s)}$ corresponding to time instants $t=1, ..., T$. Consider two hypotheses $H_i$, $i \in \{0, 1\}$ corresponding to the presence and absence of an event $E$ in the feature stream. Under each hypothesis we assume that the feature stream is generated by a hidden Markov model [26] (the parameters of the HMM under each hypothesis are learned using the EM algorithm [26]). Maximum likelihood detection is used to detect the underlying hypothesis for a given sequence of features observed and then the corresponding hidden state sequence is obtained by using the viterbi decoding (explained later). Once the state sequence for each feature stream is obtained, we can use these intermediate-level decisions from that feature stream [13] in a hierarchical approach.
Hierarchical models use these state sequences to do the inferencing. Fig. 7 shows the fusion model, where standard HMM is used for fusing the different modalities. However this model has may limitations - it assumes the different streams are independent for a particular concept and models a exponentially decaying distribution for a particular event, which as discussed is not true in general.

![Diagram of Hierarchical multimedia fusion using HMM.](image)

**Figure 7**: Hierarchical multimedia fusion using HMM. The media HMMs are responsible for mapping media observations to state sequences.

The Figure 8 illustrates in Hierarchical structure which uses DDIOMM. DDIOMM helps us to get around these providing a more suitable fusion architecture. It is discussed in more detail in the next section.

### 4.1 The Duration Dependent Input Output Markov Model

Consider a sequence $y$ of symbols $y_1, \ldots, y_T$ where $y_i \in \{1, \ldots, N\}$. Consider another sequence $q$ of symbols $q_1, \ldots, q_T$, where $q_i \in \{1, \ldots, M\}$. The model between the dotted lines in Figure 8 illustrates a Bayesian network involving these two sequences. Here $y = \{A, V\}$ (i.e. Cartesian product of the audio and video sequence). This network can be thought of as a mapping of the input sequence $y$ to the output
sequence $Q$. We term this network the *Input Output Markov Model*. This network is close in spirit to the input output hidden Markov model [47].

The transition in the output sequence is initially assumed to be Markovian. This leads to an exponential duration density model. Let us define $A=[A_{ijk}]$, $i,j\in[1,...,M]$, $k\in[1,...,N]$ as the map. $A_{ijk}=P(q_t=j|q_{t-1}=i,y_t=k)$ is estimated from the training data through frequency counting and tells us the probability of the current decision state given the current input symbol and the previous state. Once $A$ is estimated, we can then predict the output sequence $Q$ given the input sequence $Y$ using Viterbi decoding.

The algorithm for decoding the decision is presented below.

$$
\delta_t(j) = \max_{q} P(q_1,...,q_{t-1},q_t = j \mid y_1,...,y_t)
$$

\[ (5) \]
Chapter 2

\( Q_t \) indicates all possible decision sequences until time \( t-1 \). Then this can be recursively represented.

\[
\delta_t(j) = \max_{i=1}^{M} \delta_{t-1}(i) A_{ij} \\
\Delta_t(j) = \arg \max_{i=1}^{M} \delta_{t-1}(i) A_{ij} \\
P^* = \max_{i=1}^{M} \delta_T(i)
\]

\( P^* \) is the probability of the best sequence given the input and the parametric mapping \( A \). We can then backtrack the best sequence \( Q^* \) as follows

\[ q_T^* = \arg \max_{i=1}^{M} \delta_T(i) \]

**Backtracking**

\[ q_{T-l}^* = \Delta_{T-l+1}(q_{T-l+1}) \text{ for } l = 1: T - 1 \]

In the above algorithm, we have allowed the density of the duration of each decision state to be exponential. This may not be a valid assumption. To rectify this we now introduce the duration dependent input output Markov model (DDIOMM). Our approach in modeling duration in the DDIOMM is similar to that by Ramesh et al \[48\].

This model, in its standard form, assumes that the probability of staying in a particular state decays exponentially with time. However, most audio-visual events have finite duration not conforming to the exponential density. Ramesh et al \[48\] have shown that results can improve by specifically modeling the duration of staying in different states. In our current work, we show how one can enforce it in case of models like duration dependent IOMM.

Let us define a new mapping function \( A = [A_{ikld}] \), \( i,j \in [1,..,M], k \in [1,..,N], d \in [1,..,D] \) where \( A_{ikld} = P(q_k = j | q_{k-1} = i, d_{k-1}(i) = d, y_k = k) \). \( A \) can again be estimated from the training data by frequency counting. We now propose the algorithm to estimate the best decision sequence given \( A \) and the input sequence \( y \). Let

\[
\delta_i(j,d) = \max_{q} P(q_1,..,q_{t-1},q_t = j, d, (j) = d | y_1,..,y_t) \tag{6}
\]

where \( Q_t \) indicates all possible decision sequences until time \( t-1 \) and \( d(i) \) indicates the duration in terms of discreet time samples for which the path has continued to be in the state \( q_i \). This can be recursively computed as follows

\[ \delta_i(j,1) = P(q_1 | y_1) \]
\[ \delta_i(j,d) = 0, \ d > 1 \]
\[ \delta_{t+1}(j,1) = \max_{i=1}^{M} \max_{d=1}^{D} \delta_i(i,d) A_{ij} \]
\[ \delta_{t+1}(j,d+1) = \delta_i(j,d) A_{ij} \]
Let
\[
\Delta_i(j,i) = \arg \max_{d=1:D} \delta_{i-1}(i,d) A_{ijd} \quad 1 \leq i \leq j, \quad i \neq j
\]
\[
\Psi_j(j) = \arg \max_{i=1:M} \delta_{i-1}(i,\Delta_i(j,i)) A_{ij\Delta_j(j,i)}
\]
Finally
\[
\eta(i) = \arg \max_{d=1:D} \delta_T(i,d) \quad 1 \leq i \leq M
\]
\[
P^* = \max_{i=1:M} \delta_T(i,\eta(i))
\]
\(P^*\) is the probability of the best sequence given the input and the parametric mapping \(A\). We can then backtrack the best sequence \(Q^*\) as follows
\[
q_T^* = \arg \max_{i=1:M} \delta_T(i,\eta(i))
\]
\text{Backtracking}
\[
x = \eta(q_T^*) \quad t = T \quad z = x
\]
\text{while} \quad t > 1
\[
q_{t-x+1}^* = q_t^* \quad l = 1, \ldots, z - 1
\]
\[
q_{t-x}^* = \Psi_{t-x+1}(q_x^*)
\]
\[
z = \Delta_{t-x+1}(q_t^*, q_{t-x}^*)
\]
\[
t = t - x \quad x = z
\]
Using the above equations, one can decode the hidden state sequence. Each state or the group of states correspond to the state of the environment or the particular semantic concept that is being modeled. In the next section, we compare the performance of these models with the traditional HMMs and show that one can get huge improvements.
We compare the performance of our proposed algorithm with the IOMM as well as with the traditional HMM with its states being interpreted as decisions. We use the domain of movies and the audio-visual event explosion for comparison. Data from a movie is digitized. We have over 10000 frames of video data and the corresponding audio data split in 9 clips. The data is labeled manually to

5. EXPERIMENTAL SETUP, FEATURES AND RESULTS

Figure 9: Some typical frames from a Video Clip.
construct the ground truth. Figure 9 show some typical frames of the video sequence.

From the visual stream we extract several features describing the color (HSV histograms, multiple order moments), structure (edge direction histogram) and texture (statistical measures of gray level co-occurrence matrices at multiple orientations) of the stream [49]. From the audio stream we extract 15 MFCC coefficients, 15 delta coefficients and 2 energy coefficients [7]. As described in earlier we train HMMs for the positive as well as the negative hypothesis for the event explosion. HMMs for audio streams and video streams are separately trained. Each HMM has 3 states corresponding (intuitively) to the beginning, middle and end state of the event. Using the pair of models for the positive and negative hypothesis we then segment each clip into two types of segments corresponding to the presence or absence of the event. Within each segment the best state sequence decoded by the Viterbi algorithm is available to us.

For the audio and video streams synchronized at the video frame rate, we can now describe each frame by a symbol from a set of distinct symbols. Let $v^h_s$ denote the number of states $v$ corresponding to hypothesis $h$ and feature stream $s$. Then the total number of distinct symbols needed to describe each audio-visual frame jointly is given by $\prod_{s=1}^{S} \sum_{h=1}^{H} v^h_s$.

![Figure 10: Classification error for the nine video clips using the leave one clip out evaluation strategy. The maximum error for the DDIOMM is the least among the maximum error of the three schemes.](image)

This forms the input sequence $\gamma$. Similarly, with $h^s$ denoting the number of hypotheses for stream $s$ the total number of distinct symbols needed to describe
the decision for each audio-visual frame jointly is given by \( \prod_{k=1}^{s} h^k \). This forms the symbols of our decision sequence \( Q \).

**Table 1:** Comparing overall classification error.

<table>
<thead>
<tr>
<th></th>
<th>HMM</th>
<th>IOMM</th>
<th>DDIOMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Error (%)</td>
<td>20.15</td>
<td>18.52</td>
<td>13.23</td>
</tr>
</tbody>
</table>

We report results using a leave-one-clip-out strategy. The quantum of time is a single video frame. To report performance objectively, we compare the prediction of the fusion algorithm for each video frame to our ground truth. Any difference between the two constitutes to a false alarm or mis-detection. We also compare the classification error of the three schemes. Figure 10 shows error for each clip using the three schemes. Amongst the three schemes, the maximum error across all clips is least for the DDIOMM. Table 1 shows the overall classification error across all the clips.

Clearly the overall classification error is the least for the DDIOMM. We also compare the detection and false alarm rates of the three schemes. Figure 11 shows the detection and false alarm for the three schemes. Figures 10 and 11 and Table 1 thus show that the DDIOMM performs better event detection than the IOMM as well as the HMM.

**Figure 11:** Comparing detection and false alarms. DDIOMM results in best detection performance.
6. CONCLUSION

In this paper we have analyzed the problem of detecting temporal events in videos using multiple sources of information. We have argued that some of the main characteristics of this problem is the duration of the events and the dependence between the different concepts. In past, using standard HMMs for fusion, both of these issues were ignored. We have shown as how one can use the standard probabilistic models, modify them and obtain superior performance.

In particular, we present a new model, the duration dependent input output Markov model (DDIOMM) for performing integration of intermediate decisions from different feature streams to detect events in multimedia. The model provides a hierarchical mechanism to map media features to output decision sequences through intermediate state sequences. It forces the multimodal input streams to be dependent given the target event. It also supports discreet non-exponential duration models for events. By combining these two features in the framework of generative models for inference, we present a simple and efficient decision sequence decoding Viterbi algorithm. We demonstrate the strength of our model by experimenting with audio-visual data from movies and the audio-visual event explosion. Experiments comparing the DDIOMM with the IOMM as well as the HMM reveal that the DDIOMM results in lower classification error and improves detection.

REFERENCES


