Topic Transition Detection Using Hierarchical Hidden Markov and Semi-Markov Models

Dinh Q. Phung, T.V. Duong, S.Venkatesh
Department of Computing
Curtin University of Technology
Perth, Western Australia
{phungquo,duong,svetha}@cs.curtin.edu.au

Hung H. Bui
Artificial Intelligence Center
SRI International
Menlo Park, CA 94025, USA
bui@ai.sri.com

ABSTRACT

In this paper we introduce a probabilistic framework to exploit hierarchy, structure sharing and duration information for topic transition detection in videos. Our probabilistic detection framework is a combination of a shot classification step and a detection phase using hierarchical probabilistic models. We consider two models in this paper: the extended Hierarchical Hidden Markov Model (HHMM) and the Coxian Switching Hidden semi-Markov Model (S-HSMM) because they allow the natural decomposition of semantics in videos, including shared structures, to be modeled directly, and thus enabling efficient inference and reducing the sample complexity in learning. Additionally, the S-HSMM allows the duration information to be incorporated, consequently the modeling of long-term dependencies in videos is enriched through both hierarchical and duration modeling. Furthermore, the use of the Coxian distribution in the S-HSMM makes it tractable to deal with long sequences in video. Our experimentation of the proposed framework on twelve educational and training videos shows that both models outperform the baseline cases (flat HMM and HSMM) and performances reported in earlier work in topic detection. The superior performance of the S-HSMM over the HHMM verifies our belief that duration information is an important factor in video content modeling.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing.

General Terms: Algorithms, Management.

Keywords: Topic Transition Detection, Hierarchical Markov (Semi-Markov) Models, Coxian, Educational Videos.

1. INTRODUCTION

The ultimate goal of the video segmentation problem is to characterize the temporal dynamics of the video whereby it can be segmented into coherent units, possibly at different levels of abstraction. Seeking abstract units to move beyond the shots has thus been an active topic of much recent research. While the problem of shot transition is largely solved at a satisfactory level [7], the 'abstract units' or scene detection problem is much harder, partially due to the following three challenges identified in [29]: (a) the variety in directional styles, (b) the semantic relationship of neighbouring scenes, and (c) the knowledge of the viewer about the world. While the last aspect is beyond the scope of this work, the first two clearly imply that effective modeling of high-level semantics requires the domain knowledge (directional style) and the modeling of long-term, multiple-scale correlations of the video dynamics (neighboring semantic relationship).

Modeling temporal correlations over a long period is generally a challenging problem. As we shall review in the subsequent section, this problem is usually solved in a specific domain setting so that the expert knowledge about the domain can be utilised. While organization of content in generic videos (e.g., movies) is too diverse to be fully characterized by statistical models, the hierarchy of semantic structure in the class of education-oriented videos is more defined, exposing strong temporal correlation in time, and thus make it more desirable to probabilistic modeling. In this paper, we concentrate on this video genre and develop an effective framework to segment these videos into topically correlated units. This problem is an important step to enable abstraction, summarization, and browsing of educational content – a rich class of film genre that has an increasing important role in building e-services for learning and training.

Probabilistic modeling of temporal correlations in video data is however a difficult problem. It is complicated because the underlying semantics naturally possess a hierarchical decomposition with possible existence of tight structure sharing between high-level semantics. In addition, the typical duration for these structures usually varies for each of its higher semantic. As an example, assisted narration – a section that involves the narrator talking to the audience – is usually used in both the introduction and in the main body of a topic in an educational video. However while one, or rarely two, shots of assisted narration (AN) are considered sufficient for the introduction, the body typically requires many AN shots. Thus it is important to exploit and fuse hierarchical decomposition, structure sharing and duration information in a unified framework to effectively address the problem of topic transition detection.

The most widely used pobabilistic model is the hidden Markov model (HMM). However, in many cases, the HMM is unsuitable for video analysis since the strong Markov assumption makes it too restrictive to capture correlations

MM'05, November 6–11, 2005, Singapore.
Copyright 2005 ACM 1-59593-044-2/05/0011 ...$5.00.

over long periods. This limitation is usually overcome in the literature by the use of a series of HMMs in a hierarchic manner. The underlying problem in these approaches still is the manual combination of HMMs at the higher levels which results in the excessive expense of preparing the training data and, more importantly, the interaction across higher semantic levels is not incorporated during model training. One rigorous approach to overcome this limitation is the use of the Hierarchical Hidden Markov Model (HHMM), first introduced in [6] and later extended to handle structure sharing in [3]. The sophisticated model in [3] allows natural hierarchical organization of the videos, including any existing structure sharing, to be modeled rigorously. Practically this will result in computational savings and a reduction in sample complexity for learning. Given its advantages, we use this model in this paper to model educational video content for topic transition detection.

It is natural to see that durative properties play an important role in human perception. An excessively long lecture would bore the students. As such, education-oriented videos (e.g., news, documentaries, lectures, training videos, etc.) exhibit strong duration information in their content. We thus propose an alternative approach towards handling temporal dependencies over long periods through the explicit modeling of duration information captured in semi-Markov models. In these models, a state is assumed to remain unchanged for some duration of time before it transits to a new state, and thus it addresses the violation of the strong Markov assumption from having states whose duration distributions are non-geometric.

Existing semi-Markov models commonly model duration distributions as multinomials. Video data is however typically very long, thus making a multinomial semi-Markov model unsuitable for video analysis since it would result in both excessive computation and the number of parameters required. Continuous modeling of duration does exist such as in the use of the Gamma distribution, or more generally the exponential family, described in [12, 16] to provide more compact parameterization. However, there are still two limitations applied to these models for video analysis: (a) learning these distributions requires numerical optimization and the time complexity still depends on the maximum duration length, and (b) no hierarchical modeling has been attempted. Fortunately, in [5], a Switching Hidden Semi-Markov Model (S-HSMM) is introduced in which the duration is modeled as a discrete M-phase Coxian distribution. This model is particularly interesting for video analysis since: (1) it can model hierarchical decomposition, and (2) the Coxian duration modeling results in fast learning and inference, the number of parameters is small and close-formed estimation exists. Parameterizing long-term temporal correlations existing in video is thus enriched by both the hierarchical architecture and the duration modeling at the bottom level of the S-HSMM.

To model video content, we argue that it is beneficial to exploit both the hierarchical organization of the videos, their semantically shared substructures and typical durations of important semantics. These aspects are all addressed in this paper in a unified and coherent probabilistic framework. We use the HHMM and the S-HSMM and propose a two-phase architecture for detecting topical transition in educational videos. In the first phase, shots are classified into meaningful labels. Using classified shot labels, the second phase trains a hierarchical probabilistic model (HHMM or S-HSMM) which is then used at a later stage for segmentation and annotation. Prior knowledge about the domain, including shared structures, is incorporated into the topological structure during training.

Our cross-validation on a dataset including a mix of twelve videos demonstrates promising results. The performances from the baseline cases (HMM and HSMM) have shown that they are too restrictive and unsuitable in our detection scheme, proving the validity of hierarchical modeling. The performances of the hierarchical models, including the HHMM and S-HSMM, are shown to surpass all results reported in earlier work in topic detection [23, 20, 4]. The superior performance of the S-HSMM over the HHMM has also demonstrated our belief that duration information is indeed an important element in the segmentation problem.

Exploiting the hierarchy, structure sharing and duration in a unified probabilistic framework, our contributions are twofold: (1) we provide a coherent hierarchical probabilistic framework for topic detection. Although the current report concentrates on the educational genre, this framework can clearly generalize to other genres such as news and documentaries, and (2) to our knowledge we are the first to investigate duration and hierarchical modeling for video segmentation[1] in a unified framework.

The remainder of this paper is organized as follows. In the next section, we provide related background to this work. This is followed by a detailed section on the detection framework including the description of the HHMM and S-HSMM. We detail the shot classification phase in Section 4. Experimental results are then reported in Section 5. Finally, the conclusion follows in Section 6.

2. RELATED BACKGROUND

Seeking high-level semantics to move beyond the shots has been the central theme of much recent research. Attempts towards this problem have resulted in a fast growing body of work, and depending on the investigating domain, the abstracting units appear under different names such as *scene, story, episode* for motion pictures; *topic, subtopic, macro segments, story units* for information-oriented videos (news, documentaries, training and educational videos), or general term like *logical story units* used in [8, 32]. Otherwise stated, we shall the term 'scene' in this section to mean all of the aforementioned names.

Early attempts have targeted extracting scene-level concepts in broadcast programs, in particular news videos (e.g., [9, 14, 26]). In these attempts, the semantic extraction problem is usually cast as the classification problem. The authors in [26], for example, combine a number of visual and aural low-level features together with shot syntax presented in news videos to group shots into different narrative structures and label them as anchor-shot, voice-over, or inter-

[1] Since topic change coincides with a shot transition, the shot boundary provides crucial information in detecting topic transitions, therefore the term 'duration' in this work is calculated in terms of the number of shots. This drastically simplifies the modeling process. An alternative way of modeling duration is to uniformly replicate a shot label based on its length. However, doing this would require an extra modeling of shot transition knowledge. In this work, we avoid this complication and concentrate on duration information based on the shot counts.

view. Liu *et al.* [14] propose a video/audio fusion approach to segment news reports from other categories in broadcast programs with different types of classifiers (simple threshold method, Gaussian mixture classifier, and support vector machine). Ide *et al.* [9] propose an automatic indexing scheme for news video where shots are indexed based on the image content and keywords into five categories: speech/report, anchor, walking, gathering, and computer graphics. Caption text information is then used with classified shots to build the indices.

Segmentation of the *news story* is the second major theme explored in the broadcast domain. The common underlying approach used in these works is the use of explicit 'rules' about the structure of news programs to locate the transitions of a news story. Commonly accepted heuristics are for example: a news story often starts and finishes with anchor-person shots [31]; the start of a news story is usually coupled with music [2]; or a relative long silence period is the indication of the boundary between two news stories [33]. More complicated rules via temporal analysis are also exploited such as the work of [37] which utilises detection results of anchor-persons and captions to form a richer set of rules (i.e., if the same text caption appears in two consecutive anchor-person shots, then they belong to the same news story). There is also a body of work which casts the segmentation problem of news story in a HMM framework [10, 4]. The authors in [10], for example, propose the news segmentation problem as problem of decoding the maximum state sequence of a trained HMM whose transition matrix is tailored by explicit rules about the news program. A somewhat similar approach to the work in this paper is [4] (whose results came first in the TRECVID2003 story segmentation benchmark). Shots in [4] are first classified into a set common labels in news (e.g., anchor, 2anchor, text-scene, etc.). These labels are then input to a HMM for the segmentation task. They report best performances of 74.9% recall and 80.2% precision for the TRECVID dataset. The work of [4] however remains limited due to the flat structure HMM, and it is not clear how the set of 'transition' states were chosen. In an effort to move beyond flat structure, the authors of [4] have raised the need for high-order statistical techniques, which will be addressed in this paper through the HHMM and S-HSMM.

More recent approaches towards scene extraction have shifted to motion pictures (e.g., [30, 34, 1, 31]). Detecting scenes in motion pictures is in general a challenging problem and there are three main existing approaches as outlined in [31]: temporal clustering-based, rule-based and memory-based detection. In the *clustering-based* approach, shots are grouped into scenes based on visual similarity and temporal closeness (e.g., [8, 13]). Scene breaks in the *rule-based* detection approach are determined based on the semantic and syntactic analysis of audiovisual characteristics and in some cases further enhanced with more rigorous grammars from film theory (e.g., [34, 1]). The authors in [30] propose a *memory-based* scene detection framework. Visual shot similarity in these works is determined based on the consistency in color chromaticality, and the soundtrack is partitioned into 'audio scenes'. Visual and aural data are then fused within a framework of memory and attention span model to find likely scene breaks or singleton events. Further related background on scene detection can be found in many good surveys (e.g., [30, 28, 31]).

Existing HMM-based approaches towards modeling long-term temporal dependencies typically use pre-segmented training data at multiple levels, and hierarchically train a pool of HMMs, in which HMMs at the lower levels are used as input to the HMMs at the upper levels. In principle, some fundamental units are recognised by a sequence of HMMs, and then likelihood values (or labels) obtained from these HMMs are combined to form a hierarchy of HMMs[2] to capture the interactions at higher semantic levels (e.g., [11, 18]). Analysing sports videos, Kijak *et al.* [11] propose a two-tiered classification of tennis videos using two layers of HMMs. At the bottom level, four HMMs are used to model four shot classes ('first missed serve','rally', 'replay', and 'break'). Each HMM is trained separately and subsequently topped up by another HMM at the top level which represents the syntax of the tennis video with three states of the game: 'sets', 'games', and 'points'. Parameters for the top HMM are, however, all manually specified. In [18], a generic two-level hierarchy of HMMs is proposed to detect recurrent events in movies and talk shows. Their idea is to use an ergodic HMM at the top level, in which each state is another (non-ergodic) sub-HMM representing a type of signal stationary properties. For the case of movies, the top HMM has six states, and each is in turn another three-state non-ergodic HMM. The observations are modelled as a mixture of Gaussians. After training, the authors claim that interesting events can be detected such as 'explosion', 'male speech', and so on. While being able to overcome the limitation of the flat HMM in modeling long-term dependencies, approaches that use HMMs at multiple levels still suffer from two major problems: (1) pre-segmented and annotated data are needed at all levels for training, and (2) in most existing work parameterization at higher levels has to be manually specified. In many cases, preparing training data at multiple levels is extremely tedious and at worst, may not be possible. With respect to the second problem, since each semantic level has to be modeled separately, the underlying problem is that the interactions across semantic layers are not modeled and thus do not contribute to the learning process.

One framework that integrates the semantics across layers is the Hierarchical Hidden Markov Model (HHMM) proposed recently in [6]. The hierarchical HMM extends the standard HMM in a hierarchic manner to allow each state to be recursively generalised as another sub-HMM, and thus enabling the ability to handle hierarchical modeling of complex dynamic processes, in particular "the ability to infer correlated observations over long periods in the observation sequence via the higher levels of hierarchy" [6]. The original motivation in [6] was to seek better modeling of different stochastic levels and length scales presented in language (e.g., speech, handwriting, or text). However, the model introduced in [6] considers only state hierarchies that have tree structures, disallowing the sharing of substructures among the high-level states. Recognizing this need, the authors in [3] have extended the strict tree-form topology in the original HHMMs of [6] and allowed it to be a general lattice structure. The extension thus permits a state at any arbitrary level of the HHMMs to be shared by more than one parental state at its higher level (i.e., resulting in a compact form of parameter typing at multiple levels). This extended

[2] Not to be confused with the Hierarchical HMMs.

form is very attractive for video content modeling since it allows the natural organization of the video content to be modeled not only in terms of multiple scales but also in terms of shared substructures existing in the decomposition. Further details on the HHMM are provided in Section 3.1.

Early application of the HHMM for video analysis is found in [36] and later extended in [35]. In these works, the authors use the HHMM to detect the events of 'play' and 'break' in soccer videos. For inference and learning, the HHMM is 'collapsed' into a flat HMM with a very large product state space, which can then be used in conjunction with the standard forward/backward passes as in a normal HMM. Four methods are compared in [36] to detect 'play' and 'break': (1) supervised HMMs, in which each category is trained with a separate HMM, (2) supervised HHMMs, in which bottom level HMMs are learned separately and parameters for the upper levels are manually specified, (3) unsupervised HHMMs without model adaptation, and (4) supervised HHMMs with model adaptation. In (3) and (4), two-level HHMMs are used. Their results have shown a very close match between unsupervised and supervised methods in which the completely unsupervised method with model adaptation performs marginally better. These figures are $75.5\%, 75.0\%, 75.0\%$ and 75.7% respectively for those four methods. While presenting a novel contribution to the feature selection and model selection procedure, the application of the HHMMs in this work is still limited both for learning and for exploitation of the hierarchical structure. Flattening a HHMM into a flat HMM as done in [36, 35] suffers from many drawbacks as criticized in [17]: (a) it cannot provide multi-scale interpretation, (b) it loses modularity since the parameters for the flat HMM get constructed in a complex manner, and (c) it may introduce more parameters, and most importantly it does not have the ability to reuse parameters, in other words parameters for the shared sub-models are not 'tied' during the learning, but have to be replicated and thus lose the inherent strength of hierarchical modeling.

Being able to model shared structures, the extended HHMMs of [3] allows us to build more compact models, which facilitates more efficient inference and reduces the sample complexity in learning. This model is applied in [20] and [22] for the problem of topic transition detection and video structure discovery respectively. The authors in [20] use a three-level HHMM for the detection of topic transitions in educational videos. Differing from our experiments in this paper, the HHMM in [20] is modified to operate directly with continuous-valued observed data via the use of Gaussian mixture models as the emission probabilities. Each shot-based observed vector consists of seven features extracted from visual and audio streams. They report a 77.3% recall rate and 70.7% precision for the detection task. In another application, with the help of prior knowledge about educational videos, a topology for a three-level HHMM is used in [22] to automatically discover meaningful narrative units in the educational genre. Their experiments have shown encouraging results in which many meaningful structures are hierarchically discovered such as 'on-screen narration with texts', 'expressive linkage', 'expressive voice-over', etc. The work of [22] is somewhat similar to that of [18] reviewed earlier in this section, except the model in [22] allows more domain knowledge to be encoded and the parameters are all learned automatically.

3. THE PROBABILISTIC TOPIC DETECTION FRAMEWORK

Our topic detection framework consists of two phases. The first phase performs shot detection and low level feature extraction and then classifies a shot in a meaningful label set Σ. This phase is described in Section 4. In the next phase, we train a HHMM or S-HSMM over the alphabet space Σ from the training data and then use it in conjunction with the Viterbi to perform segmentation and annotation. The architecture of the framework is depicted in Figure-1.

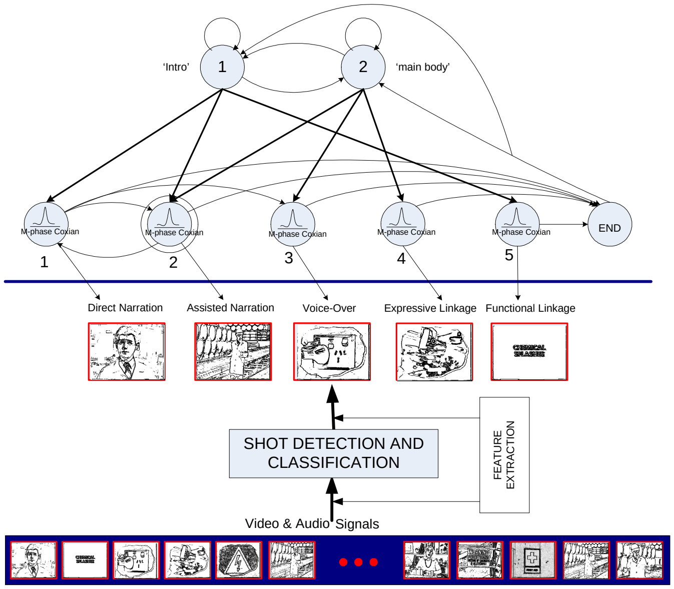

Figure 1: The architecture for topic detection framework.

The two-level HHMM and the S-HSMM (whose topology is shown on the top of Figure-1) are special cases of the hierarchical model with two layers. For the S-HSMM (HHMM), the top layer is a Markov sequence of *switching* variables, while the bottom layer is a sequence of concatenated HSMMs (HMMs) whose parameters are determined by the switching variables at the top. Thus, the dynamics and duration parameters of the HSMM (HMM) at the bottom layer are not time invariant, but are 'switched' from time to time, similar to the way the linear Gaussian dynamics are switched in a switching Kalman filter. When mapping to the topic modeling problem, the bottom layer is used to capture 'atomic' semantics such as voice-over, expressive linkage or assisted narration. Combinations of these atomic semantics then form higher-level semantics, each of which is represented by a hidden state at the top layer in our model.

3.1 The Hierarchical HMM

With the assumed knowledge of the flat HMM (e.g., see [24]), we shall now briefly describe the HHMMs. A hierarchical HMM is formally defined by a three-turple $\langle \zeta, \theta, \Sigma \rangle$: a topological structure ζ parameterized by θ and an emission alphabet space Σ. The topology ζ specifies the model depth D, the state space $\mathcal{S}^d$ available at each level d, and the parent-children relationship between two consecutive levels. For example, the two-level topology shown on the top of

Figure-1 specifies the children set at the bottom level for state 1 is $\{1, 2, 5\}$ and for state 2 is $\{2, 3, 4\}$. Here, state 2 at the bottom level has been 'shared' by both state 1 and 2 at the top level. Given ζ, the parameter θ of the HHMM is specified in the following way. For $d < D$, $p \in \mathcal{S}^d$ and $i, j \in \mathcal{S}^{d+1}$ are the children of p: $\pi_i^{d,p}$ is the initial probability of i given p; $A_{i,j}^{d,p}$ is the transition probability from i to j given the parent p; and $A_{i,end}^{d,p}$ is the probability that state i going to end-state (i.e., returns control to its parent) given the parent is p. Finally, for each state i at the lowest level D and an alphabet $v \in \Sigma$: $B_{v|i}$ is the emission probability of observing v given the current state at the lowest level is i. The whole parameter set is written compactly as: $\theta = \{\pi, A, A_{end}, B\}$, where:

$$\pi = \bigcup_{\substack{1 \le d < D \\ p \in \mathcal{S}^d}} \left\{ \pi^{d,p} : 1 \times M \right\}, \quad B : |\mathcal{S}^d| \times |\Sigma|$$

$$A = \bigcup_{\substack{1 \le d < D \\ p \in \mathcal{S}^d}} \left\{ A^{d,p} : M \times M \right\}, \; A_{end} = \bigcup_{\substack{1 \le d < D \\ p \in \mathcal{S}^d}} \left\{ A_{end}^{d,p} : 1 \times M \right\}$$

where in each each M is implicitly meant the number of children of p and $|.|$ is the cardinality operator. Stochastic constraints require: $\sum_i \pi_i^{d,p} = 1, \sum_v B_{v|i} = 1$ and $\sum_j A_{i,j}^{d,p} + A_{i,end}^{d,p} = 1$. An intuitive way to view the set θ is to consider the subset $\{\pi^{d,p}, A^{d,p}, A_{end}^{d,p}\}$ as the parameter of the p-initiated Markov chain at level d. This chain is terminated when one of the children i of p reaches the end-state with the probability of $A_{i,end}^{d,p}$. For inference and learning, the HHMM is represented as a dynamic Bayesian network (DBN) and can be learned by the Asymmetric Inside-Outside algorithm in [3] or by the forward/backward passes in [17]. Figure-3 shows on its left the DBN representation of the HHMM with two levels, i.e., $D = 2$. We refer readers to [6, 17, 3] for further information on the HHMMs.

3.2 The Switching-Hidden Semi Markov Model

To provide an intuitive view to the S-HSMM, starting from the description of the HHMMs from the previous section, let us consider the case of a two-layer HHMM ($D = 2$) defined as follows. The state space is divided into the set of states at the top level $Q^* = \mathcal{S}^1 = \{1, \ldots, |Q^*|\}$ and states at the bottom level $Q = \mathcal{S}^2 = \{1, \ldots, |Q|\}$. This model is parameterized by $\theta = \{\pi^*, A^*, \pi, A, A_{end}, B\}$.

At the top level, π_p^* and A_{pq}^* are respectively the initial probability and the transition matrix of a Markov chain defined over the state space Q^*. For each state $p \in Q^*$, $\mathrm{ch}(p) \in Q$ is used to denote the set of children of p. As in the case of the extended HHMM in [3], it is possible that different parent states may share certain common children, i.e., $\mathrm{ch}(p) \cap \mathrm{ch}(q) \neq \emptyset$ for $p, q \in Q^*$. A transition to p at the top level Markov chain will initiate a sub-Markov chain at the lower level over the state space $\mathrm{ch}(p)$ parameterized by $\{\pi^p, A^p, A_{end}^p\}$ where π_i^q and A_{ij}^p are the initial and transition probabilities as in the normal HMM setting, and $A_{i,end}^p$ is the probability that this chain will terminate after a transition to i. At each time point t, a discrete symbol $y_t \in \Sigma$ is generated with a probability of $B_{v|i}$ where i is the current state at the bottom level. In the description of this two-level HHMM, the duration d for which a bottom state i remains the same clearly has a geometric distribution parameterized by its non-self-transition probability $(1 - A_{ii}^p)$, i.e., $d \sim \mathrm{Geom}(1 - A_{ii}^p)$.

In many cases, the geometric distributions are often too restricted to model realistic data. The Switching Hidden Semi-Markov Models (S-HSMMs) proposed in [5] overcomes this restriction and allows the duration d of state i at the bottom level to follow a more general discrete distribution $d \sim D_d^{p,i}$. More precisely, the p-initiated chain at the bottom level is now a semi-Markov sequence parameterized by $\{\pi_i^p, A_{ij}^p, D_d^{p,i}\}$ as opposed to the normal Markov chain in the HHMM case. The authors in [5] consider two families of distributions for modeling the duration: the multinomial and the Coxian. However for the multinomial case, the complexity of the learning algorithm is proportional to the maximum duration length, thus making it unsuitable for the problem of modeling video data which is usually very long in nature. Apart from the disadvantage of assuming a maximum duration, our empirical testing on the multinomial case with the maximum length of 50 has also shown that it is about 20 times slower than its Coxian counterpart reported in this paper, thus making it impractical in our settings. We will therefore omit the multinomial case and will consider exclusively the Coxian parameterization in this paper.

A discrete M-phase Coxian distribution $\mathrm{Cox}(\boldsymbol{\mu}; \boldsymbol{\lambda})$, parameterized by $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_M\}$ ($\sum_i \mu_i = 1$) and $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_M\}$, is defined as a mixture of $\left(\sum_{i=1}^M \mu_i S_i\right)$ where $S_i \triangleq (X_i + \ldots + X_M)$, in which X_i are independent random variables having geometric distributions $X_i \sim \mathrm{Geom}(\lambda_i)$. This distribution is a member of the phase-type distribution family and has the following very appealing interpretation. Let us construct a Markov chain with $M + 1$ states numbered sequentially with the self transition parameter $A_{ii} = 1 - \lambda_i$ as shown in Figure-2. The first M states represent M phases, while the last is the absorbing state which acts like an end state. The duration of each individual state (phase) i is $X_i \sim \mathrm{Geom}(\lambda_i)$. If we start from state i, the duration of Markov chain before the end state reached is $S_i = X_i + \ldots + X_M$. Thus, $\mathrm{Cox}(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is indeed the distribution of the duration of this constructed Markov chain with $\boldsymbol{\mu}$ as the initial state (phase) distribution. The discrete Coxian is much more flexible than the geometric distribution: its probability mass function is no longer monotonically decreasing and it can have more than one mode.

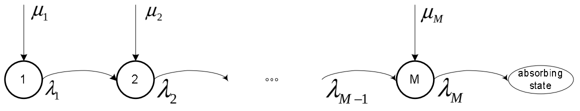

Figure 2: The phase diagram of an M-phase Coxian.

Using the Coxian distribution, the duration for the states at the bottom level in the S-HSMM is modeled as follows. For each p-initiated semi-Markov sequence, the duration of a child state i is distributed according to $D_d^{p,i} = \mathrm{Cox}(d; \boldsymbol{\mu}^{p,i}, \boldsymbol{\lambda}^{p,i})$. The parameter $\boldsymbol{\mu}^{p,i}$ and $\boldsymbol{\lambda}^{p,i}$ are M-dimensional vectors where M is a fixed number representing the number of phases in the discrete Coxian. It is easy to verify that for $M = 1$, the model reduces identically to a two-layer HHMM.

3.3 Inference and Learning in the S-HSMM

For inference and learning, the S-HSMM is represented as a dynamic Bayesian network as shown in Figure-3 and then forward/backward passes are applied to compute the filtering and smoothing distributions required for EM learning.

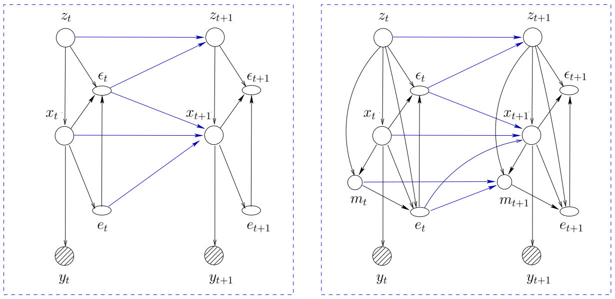

Figure 3: Two-slice DBN representation of a two-level HHMM (left) and the (Coxian) S-HSMM (right).

At each time-slice t, an amalgamated hidden state $S_t = \{z_t, \epsilon_t, x_t, e_t, m_t\}$ together with the observation y_t are maintained. The top level state is updated via z_t and ϵ_t is a boolean-valued variable set to 1 when the z_t-initiated semi-Markov sequence ends at t. At the bottom level, x_t is the current child state in the z_t-initiated chain, m_t represents the current phase of x_t and e_t is a boolean-valued variable set to 1 when x_t reaches the end of its duration. The forward and backward procedures in the general DBN are then used to compute the filtering distribution $\Pr(S_t|y_{1:t})$ and two smoothing distributions $\Pr(S_t|y_{1:T})$ and $\Pr(S_t, S_{t+1}|y_{1:T})$. With these smoothing distributions, it is sufficient to derive all expected sufficient statistics required during EM learning. The overall complexity for the forward pass (and also for the EM) is $O(|Q|^2|Q^*|^2MT)$. Further information can be found in [5].

3.4 Viterbi decoding for segmentation

To compute the best sequence state, that is to find:

$$S^*_{1:T} = \underset{S_{1:T}}{\operatorname{argmax}} \Pr(S_{1:T}|y_{1:T})$$

Viterbi decoding algorithms for the HHMM and S-HSMM are developed. These algorithms are similar to the one used in the standard HMM outlined in [24] except we replace the normal state in the HMM setting by our amalgamated state S_t which $\triangleq \{z_t, x_t, \epsilon_t, m_t, e_t\}$ for the S-HSMM and $\triangleq \{z_t, x_t, \epsilon_t, e_t\}$ for the HHMM (cf. Figure-3).

4. SHOT-BASED SEMANTIC CLASSIFICATION

In this section, we detail the first phase in the detection framework. This includes the formulation of an alphabet set Σ for shot labeling, low-level feature extraction and shot classification.

4.1 Shot labels set: Σ

Existing work on the educational videos analysis (e.g., [21, 19]) has studied the nature of this genre carefully. As noted in [21], the axiomatic distinction of the educational genre is in its purpose of *teaching* and *training*; and as such a well-crafted segment that moves viewers to actions or retains a long-lasting message requires elaborative directing skills[3]. Based on a narrative analysis used in the educational domain and observed rules and conventions in the production of this media, the authors in [21] propose a hierarchy of narrative structures at the shot level as shown in Figure-4.

In this paper, we select the five most meaningful structures from this hierarchy for experimentation. This set Σ includes: *direct-narration* (DN), *assisted-narration* (AN), *voice-over* (VO), *expressive-linkage* (EL), and *functional-linkage* (FL). We shall now briefly describe these narratives.

Direct-narration (DN) and assisted-narration (AN) are referred to jointly as on-screen narration, which refer to the segments with the appearance of the narrator. The purpose of these sections is to speak to the viewers with the voice of authority, and is commonly used to demarcate a new topic or subtopic, to clarify a concept or to lead the viewers through a procedure with examples. DN is a more strict form of on-screen narration. It involves eye-to-eye contact where the narrator speaks to the viewers directly. An analogy from news video is the anchor-shot. AN refers to parts of the video when a narrator appears in a more diverse style, and the attention of the viewers is not necessarily focused on him or her. Here, the purpose is not only to talk to the viewers, but also to emphasize a message by means of text captions and/or to convey an experience via background scenes. A similar structure from news for AN is the reporting shot. Assisted narration can be used both in the introduction of a topic or in the main body, and thus this structure should be shared[4] by both higher semantics 'introduction' and 'main body'. As we see later, this knowledge is explicitly modeled and incorporated in the design of the topology for the S-HSMM. An important feature is that although the semantics of AN is shared, the typical durations are different when it is used in the introduction or the main body respectively. An AN section used to demarcate a new topic usually contains only one, and sometimes two shots, while an AN section used in the main body is typically long, spanning a number of shots. Conditioning on the parent (i.e., introduction or main body), the typical duration distribution of the AN section is learned automatically for each case by our model.

The voice-over (VO) structure is identified as sections where the audiotrack is dominated by the voice of the narrator, *but without* his or her appearance. The purpose of these segments is to communicate with the viewers via the narrator's voice. Additional pictorial illustration is usually further shown in the visual channel.

Expressive linkage (EL) and Functional linkage (FL) belong to the same broader linkage group in the hierarchy in Figure-4. The purpose of the linkage structure is to maintain the continuity of a story line but there is *neither* on-screen *nor* voice-over narration involved. Functional linkage contains transition shots encountered in switching from one subject to the next. Usually, large superimposed text captions are used and the voice narration is completely stopped

[3] We note that the two closest video genre to educational videos is news and documentaries. In the description of what follows on educational genre, we can spot several similarities across these genre.

[4] In terms of parameterization, it is a form of *parameter tying*.

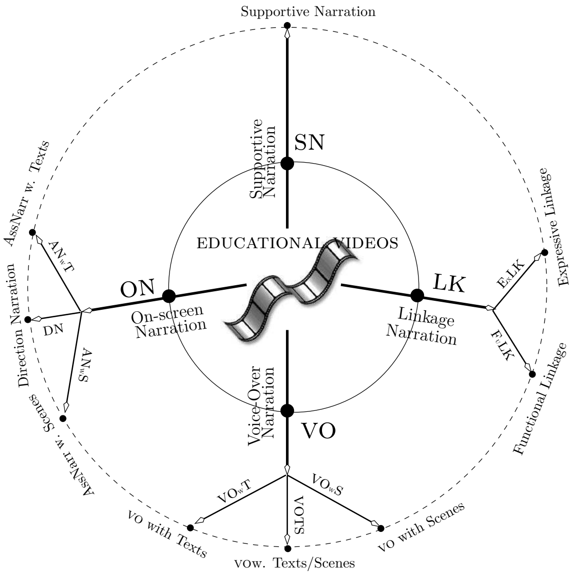

Figure 4: The hierarchy of narrative structures in educational videos proposed in [21].

with possibly music played in the background. Expressive linkage, on the other hand, is used to create 'mood' for the subject being presented. For example, in the video presenting the fire safety topic, there is a segment in which the narration is completely stopped and then a sequence of pictures of the house on fire is shown. These scenes obviously do not give any direct instruction, rather they create a sense of 'mood' that helps the video to be more appealing and interesting.

4.2 Feature extraction and shot classification

The feature set and method for shot classification described in [21] is employed in this paper. The feature set is extracted from both visual and audio streams at the shot-based level. From the image sequence, we choose to detect the frontal faces to reflect the appearance of the narrator using the CMU face detection algoritm [25]; and captioned texts as one of the common means of conveying information in educational videos using the algorithm described in [27]. In order to classify a shot into direct-narration, voice-over, linkage, etc., further information is sought from the audio stream. Audio features are computed as the percentage of the following audio classes within a shot: vocal speech, music, silence, and non-literal sound. A shot is then classified into one of the elements of $\Sigma = \{DN, AN, VO, EL, FL\}$ using the classification framework reported in [21]. Since we claim no contribution at this stage, we shall refer readers to [21] for full details on this classification scheme.

5. EXPERIMENTAL RESULTS

5.1 Data and Shot-based classification

Our dataset D consists of 12 educational and training videos containing different types of subjects and presentational styles, and thus this constitutes a relatively noisy set of data. We manually provide groundtruth for these videos with topic transitions. In some cases, the groundtruth for topic transitions comes directly from the hardcopy guidelines supplied by the producer.

At the pre-processing stage, Webflix [15] is used to perform shot transition detection and all detection errors are corrected manually. Since our contribution from this paper is at the semantic level, the latter step is to ensure an error at the shot detection does not influence the performance of the system at higher levels. Since educational videos mainly contain cut and dissolve transitions, the shot detection accuracy is found to be very high with rare cases being erroneous. Given shot indices, each video is processed as described in Section 4, and then each shot S is labeled as one of the elements of $\Sigma = \{DN, AN, VO, EL, FL\}$.

5.2 Model topology and parameterization

We will use four models in this experiments: the flat HMM and HSMM (as the baseline cases), the HHMM and the S-HSMM. For the flat HMM and HSMM, we range the number of states from 2 to 5 with the observation space Σ, where 2 is intended to be the minimum number of states required (like 'intro' and 'main body') and 5 is the number of alphabets (i.e., in the relaxed way that the number of states equates to the number of alphabets). The semi-Markov version HSMM is further parameterized by 3-phase Coxian distributions as the duration distributions of the states. The choice of $M = 3$ phases is hinted by the results reported in [5] where $M = 3$ has resulted in best performances.

For the HHMM and the S-HSMM, the topology shown in the top of Figure-1 is used to construct the S-HSMM in this experiment. This topology specifies $Q^* = 2$ states at the top level where state 1 and 2 correspond to the introduction and the main body of the topic respectively. The Markov chain at this level is similar to the flat HMM used in [4] for news story segmentation[5] reviewed in Section 2. We incorporate the assumed prior knowledge that a topic usually starts with either direct-narration, assisted-narration or functional linkage, thus state 1 has $\{1, 2, 5\}$ as its children set. Similarly, the main body can contain assisted-narration, voice-over or expressive linkage, hence its children set is $\{2, 3, 4\}$. Here state 2 (assisted narration) has been shared by both parent state 1 ('intro') and 2 ('main body'). The bottom level has 5 states corresponding to 5 shot labels. To map the labels to the bottom states, we construct a diagonal-like B observation matrix and fix it, i.e., we do not learn B. The diagonal entries of B are set to 0.99 to relax the uncertainty during the classification stage. The duration models in the S-HSMM are used with $M = 3$ phases Coxian.

5.3 Detection Results

Given the dataset D, our evaluation employs a *leave-one-out strategy* to ensure an objective cross-validation. We sequentially pick out a video V and use the remainder set $\{D \setminus V\}$ to train the model, and then use V for testing. In the results that follow, this method is used for all cases including the flat HMM, the flat HSMM, hierarchical HMM, and the S-HSMM. A topic transition is detected when the introduction state at the top level is reached during the Viterbi decoding. Examples of Viterbi decoding with the S-HSMM and HHMM are shown in Figure-5.

To measure the performance, in addition to the well-known

[5] They called 'transition' and 'internal' states instead of 'introduction' and 'main body'.

recall (RECALL) and precision (PREC) metrics, we include the F-score (F-SCORE) metric defined as:

$$\textsc{f-score} = 2 \times \frac{\textsc{recall} \times \textsc{prec}}{\textsc{recall} + \textsc{prec}} = 2 \times \left(\frac{1}{\textsc{recall}} + \frac{1}{\textsc{prec}} \right)^{-1}$$

While the recall rate measures how well the system can recover the true topic transitions, and high precision ensures that it does not over-segment the video, the F-score shows the overall performance of the system. In the ideal case when RECALL=PREC=100%, clearly F-SCORE = 1, i.e., the highest performance the system can achieve.

The baseline cases: flat HMM and HSMM
Since initialization is crucial during EM learning, we apply multiple random restart points when conducting the experiments, including the uniform initialization. Although several restarts were used, the flat HMM is found to yield extremely poor results in all cases. Even when we train and test on the same dataset, the flat HMM still produces poor detection results, proving to be unsuitable in our topical transition detection settings.

The flat HSMM produces slightly better results than the flat HMM, but still in all ten runs, the performance is still very low (RECALL= 7.74% and PREC= 48% in the best case). The poor performance of the HMM and HSMM is of no surprise, since their forms are too strict to model a rather high concept - the 'topic'. Furthermore, with the flat structures, they offer no mechanism to incorporate prior domain knowledge such as those that we use in the topology of the S-HSMM and HHMM. This clearly shows that hierarchical models are much more suitable for video analysis than the flat ones. Given the poor results in the flat structure cases, we will omit the HMM and HSMM in the discussion of what follows below.

Detection with the S-HSMM and HHMM
The recall rate, precision and F-score for representative runs are reported in Table 1, in which the best performance are highlighted in bold. The detection results for each individual video for the best cases are shown in Table 2. With different random restarting points, including the uniform initialization, the performance of the HHMM ranges from poor to very good (41.29% → 83.23% for recall and 80.00% → 84.47% for precision), whereas the S-HSMM consistently yields good results (83.87% → 84.52% for recall and 87.92% → 88.51% for precision).

Since during training there is nothing exposed to the testing examples, we also report (in the second part of Table 1) the performances of the HHMM and S-HSMM in a likelihood-based 'best model selection' scheme. This scheme works as follows. As in the leave-one-out strategy, let V be a video selected from D, and N is the number of times we train the model using the dataset $\{D \setminus V\}$ (i.e., without V). Let $\theta_i(V)$ and $\mathcal{L}_i(V)$ $(i = 1 \ldots N)$ respectively be the learned model and the likelihood (at convergence) obtained for i-th run. We then use the model θ_{i^*} to test on the unseen video V where $i^* = \underset{i=1 \ldots N}{\operatorname{argmax}} \mathcal{L}_i(V)$. Simply speaking, we sequentially 'throw away' a video V, then select the best model (i.e., highest likelihood) among all runs to test on V. For the HHMM, the result stays the same as when we choose the best performance based on the F-score. For the S-HSMM, the recall stays the same, while the precision slightly decreases from 88.51% to 87.92%. Nevertheless, the S-HSMM is still superior to the HHMM.

		RECALL (%)	PREC (%)	F-SCORE
RESULTS FOR BEST PERFORMANCE SELECTION				
HHMM	Uniform	42.58	81.48	0.559
	Rand. 1	83.23	84.47	0.840
	Rand. 2	**83.23**	**84.87**	**0.840**
	Rand. 3	83.23	84.87	0.840
	Rand. 3	41.29	80.00	0.545
	Rand. 4	83.87	83.87	0.839
S-HSMM	Uniform	84.52	87.92	0.862
	Rand. 1	**84.52**	**88.51**	**0.865**
	Rand. 2	83.87	87.25	0.855
	Rand. 3	84.52	88.51	0.865
	Rand. 4	83.87	87.25	0.855
	Rand. 5	84.52	88.51	0.865
RESULTS FOR BEST MODEL SELECTION				
HHMM		83.23	84.87	0.840
S-HSMM		84.52	87.92	0.862

Table 1: Detection Performances for the S-HSMM and the HHMM. Best performance for each case is highlighted in **bold** (we note that best performances are attained in multiple cases and we select one of them to highlight)**.**

Table 1 and 2 show that modeling with the S-HSMM results in better performances than the HHMM in both recall and precision rates. And as a result, the F-score improves from 0.840 to 0.865. While the recall rate improves only slightly, the ∼ 4% improvement in the precision indicates that the HHMM tends to over-segment the video more frequently than the S-HSMM. This has confirmed our belief that duration information is an important factor in our topic transition detection settings. The semi-Markov modeling has effectively overcome the limitation of the strict Markov assumption of $\{\text{future} \perp\!\!\!\perp \text{past} \mid \text{present}\}$[6] in the flat HMM, allowing longer temporal dependency to be captured via the duration of the state. Nevertheless, given a somewhat more contained set of data used in this experiment, the results from both the S-HSMM and HHMM are better than the previous detection results of news story reported in [4] (which came first in TRECVIC2003 testbed) and the heuristics and Bayesian approaches on topic detection in [23, 21]. These results do not only imply the advantages of the S-HSMM over the HHMM, but also show the contribution of the HHMM in its own right.

6. CONCLUSION

In this paper we explore the difficult problem of detecting topic transitions through the use of two probabilistic models, the HHMM and the S-HSMM. Both allow the modeling of hierarchy and the sharing of substructures within the hierarchy, whilst the S-HSMM additionally allows the explicit modeling of durative properties. Coupled with the use of the Coxian model, we show how this unified framework performs better than the baseline cases (the flat HMM and HSMM) and previous results reported. In particular the use of the S-HSMM demonstrates that the modeling of duration is a

[6] i.e., the future is conditionally independent of the past given the present.

Video	**TP**		**FP**		**Miss**		**GT**
1 - “EESafety”	10	8	1	3	3	5	13
2 - “SSFall”	4	4	1	1	2	2	6
3 - “ElectS”	6	6	2	1	2	2	8
4 - “TrainHaz”	18	20	2	2	3	1	21
5 - “EyeS”	10	10	0	1	0	0	10
6 - “FootS”	10	10	1	1	1	1	11
7 - “HKeeping”	11	11	3	3	1	1	12
8 - “Maintn”	9	8	1	3	4	5	13
9 - “HandS”	9	9	1	1	1	1	10
10 - “SBurning”	19	19	1	1	2	2	21
11 - “HeadProt”	6	5	1	3	1	2	7
12 - “WeldingS”	19	19	3	3	4	4	23
Sum	**131**	**129**	**17**	**23**	**24**	**26**	**155**

Table 2: Detection results for each video in the best performance cases of the S-HSMM and the HHMM (TP: True Positive, FP: False Positive, GT: Ground-Truth).

powerful tool in the extraction of higher level semantics.

The results demonstrate the promise of the approach and although the results are demonstrated with the educational and training film genre, the method can easily be applied to other genres. We believe that the promise of the approach lies in its unified probabilistic handling of durative properties and shared hierarchical structure, allowing it to handle long video sequences with inherent variability and complicated semantics.

Acknowledgement

Hung Bui is supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under Contract No. NBCHD030010.

7. REFERENCES

[1] B. Adams, C. Dorai, and S. Venkatesh. Automated film rhythm extraction for scene analysis. In *IEEE International Conference on Multimedia and Expo*, pages 1056–1059, Tokyo, Japan, August 2001.

[2] P. Aigrain, P. Jolly, and V. Longueville. Medium knowledge-based macro-segmentation of video into sequences. In M. Maybury, editor, *Intelligent Multimedia Information Retrieval*, pages 159–174. AAAI Press/MIT Press, 1998.

[3] H. H. Bui, D. Q. Phung, and S. Venkatesh. Hierarchical hidden markov models with general state hierarchy. In D. L. McGuinness and G. Ferguson, editors, *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pages 324–329, San Jose, California, USA, 2004. AAAI Press / The MIT Press.

[4] L. Chaisorn, T.-S. Chua, C.-H. Lee, and Q. Tian. A hierarchical approach to story segmentation of large broadcast news video corpus. In *IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, June 2004.

[5] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the Switching Hidden Semi-Markov Model. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 838–845, San Diego, 20-26 June 2005. IEEE Computer Society.

[6] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[7] A. Hanjalic. Shot-boundary detection: Unraveled and resolved? *IEEE Transaction in Circuits and Systems for Video Technology*, 12(2):90–105, 2002.

[8] A. Hanjalic, R. L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video retrieval systems. *IEEE Transactions in Circuits and Systems for Video Technology*, 9(4):580–588, 1999.

[9] I. Ide, K. Yamamoto, and H. Tanaka. Automatic video indexing based on shot classification. In *First International Conference on Advanced Multimedia Content Processing*, pages 99–114, Osaka, Japan, November 1998.

[10] U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll. New approaches to audio-visual segmentation of TV news for automatic topic retrieval. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 3, pages 1397–1400, Salt Lake City, Utah, 2001.

[11] E. Kijak, L. Oisel, and P. Gros. Hierarchical structure analysis of sport videos using HMMs. In *Int. Conf. on Image Processing*, volume 2, pages II–1025–8 vol.3, 2003.

[12] S. E. Levinson. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech and Language*, 1(1):2945, March 1986.

[13] T. Lin and H. J. Zhang. Automatic video scene extraction by shot grouping. *Pattern Recognition*, 4:39–42, 2000.

[14] Z. Liu and Q. Huang. Detecting news reporting using audio/visual information. In *International Conference on Image Processing*, pages 24–28, Kobe, Japan, October 1999.

[15] Mediaware-Company. Mediaware solution webflix professional V1.5.3, 1999. http://www.mediaware.com.au/webflix.html.

[16] C. D. Mitchell and L. H. Jamieson. Modeling duration in a hidden markov model with the exponential family. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages II.331–II.334, Minneapolis, Minnesota, April 1993.

[17] K. Murphy and M. Paskin. Linear-time inference in hierarchical HMMs. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, Cambridge, MA, 2001. MIT Press.

[18] M. R. Naphade and T. S. Huang. Discovering recurrent events in video using unsupervised methods. In *Int. Conf. om Image Processing*, volume 2, pages 13–16, Rochester, NY, USA, 2002.

[19] D. Q. Phung. *Probabilistic and Film Grammar Based Methods for Video Content Analysis*. PhD thesis, Curtin University of Technology, Australia, 2005.

[20] D. Q. Phung, H. H. Bui, and S. Venkatesh. Content structure discovery in educational videos with shared

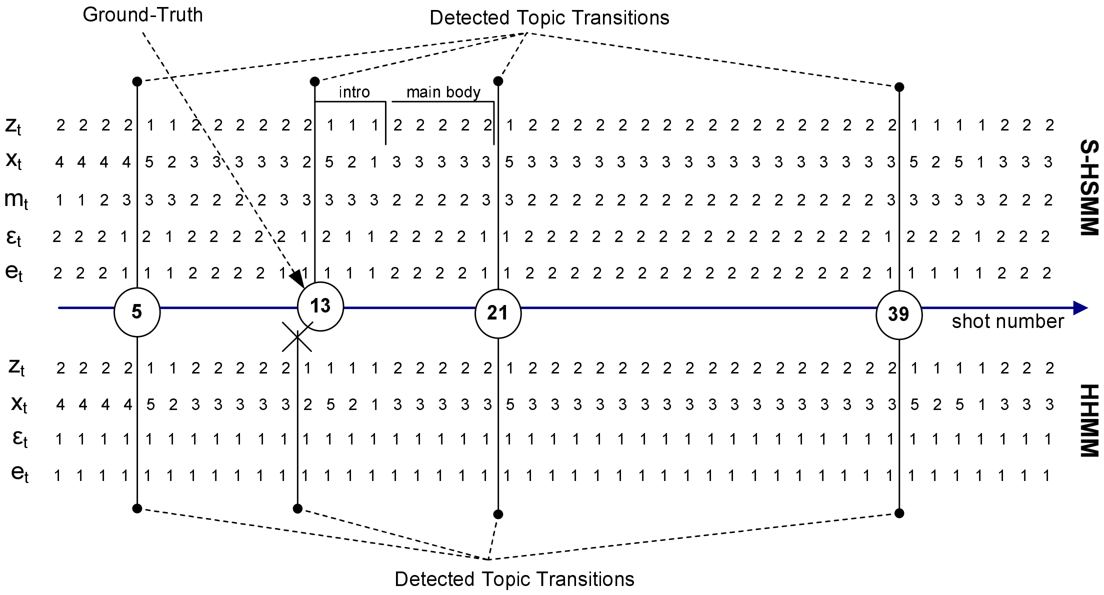
Ground-Truth
Detected Topic Transitions
intro
main body
z_t 2 2 2 2 1 1 2 2 2 2 2 2 1 1 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2
x_t 4 4 4 4 5 2 3 3 3 3 3 2 5 2 1 3 3 3 3 3 5 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 5 2 5 1 3 3 3
m_t 1 1 2 3 3 3 2 2 2 2 3 3 3 3 3 2 2 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 2 2 2
ε_t 2 2 2 1 2 1 2 2 2 2 2 1 2 1 1 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2
e_t 2 2 2 1 1 1 2 2 2 2 1 1 1 1 1 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2
S-HSMM
5
13
21
39
shot number
z_t 2 2 2 2 1 1 2 2 2 2 2 1 1 1 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2
x_t 4 4 4 4 5 2 3 3 3 3 3 2 5 2 1 3 3 3 3 3 5 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 5 2 5 1 3 3 3
ε_t 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
e_t 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
HHMM
Detected Topic Transitions

Figure 5: Example of Viterbi decoding for the S-HSMM and the HHMM for the first 45 shots of video 'EESafety'. These results should be read together with Figure-3 to see the semantics of the DBN structure.

structures in the hierarchical HMMs. In *Joint Int. Workshop on Syntactic and Structural Pattern Recognition*, pages 1155–1163, Lisbon, Portugal, August 18–20 2004.

[21] D. Q. Phung and S. Venkatesh. Structural unit identification and segmentation of topical content in educational videos. Technical report, Department of Computing, Curtin University of Technology, 2005. TR-May-2005.

[22] D. Q. Phung, S. Venkatesh, and H. H. Bui. Automatically learning structural units in educational videos using the hierarchical HMMs. In *International Conference on Image Processing*, Singapore, 2004.

[23] D. Q. Phung, S. Venkatesh, and C. Dorai. High level segmentation of instructional videos based on the content density function. In *ACM International Conference on Multimedia*, pages 295–298, Juan Les Pins, France, 1-6 December 2002.

[24] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Procs. IEEE*, volume 77, pages 257–286, February 1989.

[25] H. A. Rowley, S. Baluja, and T. Kanade. Neutral network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[26] K. Shearer, C. Dorai, and S. Venkatesh. Incorporating domain knowlege with video and voice data analysis. In *Workshop on Multimedia Data Minning*, Boston, USA, August 2000.

[27] J.-C. Shim, C. Dorai, and R. Bolle. Automatic text extraction from video for content-based annotation and retrieval. In *International Conference on Pattern Recognition*, volume 1, pages 618–620, Brisbane, Australia, August 1998.

[28] C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 2004. In Press.

[29] H. Sundaram. *Segmentation, Structure Detection and Summarization of Multimedia Sequences*. PhD thesis, Columbia University, 2002.

[30] H. Sundaram and S.-F. Chang. Computable scenes and structures in films. *IEEE Transactions in Multimedia*, 4(4):482–491, 2002.

[31] B. T. Truong. *An Investigation into Structural and Expressive Elements in Film*. PhD thesis, Curtin University of Technology, 2004.

[32] J. Vendrig and M. Worring. Systematic evaluation of logical story unit segmentation. *IEEE Transactions on Multimedia*, 4(4):492–499, 2002.

[33] C. Wang, Y. Wang, H. Liu, and Y. He. Automatic story segmentation of news video based on audio-visual features and text information. In *Int. Conf. on Machine Learning and Cybernetics*, volume 5, pages 3008–3011, 2003.

[34] J. Wang, T.-S. Chua, and L. Chen. Cinematic-based model for scene boundary detection. In *The Eight Conference on Multimedia Modeling*, Amsterdam, Netherland, 5-7 November 2001.

[35] L. Xie and S.-F. Chang. Unsupervised mining of statistical temporal structures in video. In A. Rosenfield, D. Doreman, and D. Dementhons, editors, *Video Mining*. Kluwer Academic Publishers, June 2003.

[36] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Learning hierarhical hidden markov models for unsupervised structure discovery from video. Technical report, Columbia University, 2002.

[37] X. Zhu, L. Wu, X. Xue, X. Lu, and J. Fan. Automatic scene detection in news program by integrating visual feature and rules. In *IEEE Pacific-Rim Conference on Multimedia*, pages 837–842, Beijing, China, 2001.