AUTOMATICALLY LEARNING STRUCTURAL UNITS IN EDUCATIONAL VIDEOS WITH THE HIERARCHICAL HIDDEN MARKOV MODELS

Dinh Q. Phung, Svetha Venkatesh

School of Computing
Curtin University of Technology
GPO Box U1987, Perth, 6845, Western Australia
{phungquo,svetha}@computing.edu.au

Hung H. Bui

Artificial Intelligence Center
SRI International, 333 Ravenswood Ave
Menlo Park, CA 94025, USA
bui@ai.sri.com

ABSTRACT

In this paper we present a coherent approach using the hierarchical HMM with shared structures to extract the structural units that form the building blocks of an education/training video. Rather than using hand-crafted approaches to define the structural units, we use the data from nine training videos to learn the parameters of the HHMM, and thus naturally extract the hierarchy. We then study this hierarchy and examine the nature of the structure at different levels of abstraction. Since the observable is continuous, we also show how to extend the parameter learning in the HHMM to deal with continuous observations.

1. INTRODUCTION

The indexing of multimedia content is an important issue in the management of vast amount of data. This involves the ability to automatically categorise the data at different levels of semantic abstraction by combining the cues from all modalities. Whilst this categorisation can be achieved to some extent by a bottom-up approach that uses the visual and aural media explicitly, it is our contention that higher levels of abstraction can be extracted if a data driven bottom up approach is combined with top-down contextual domain knowledge.

In this paper we examine a category of video, the education and training videos, whose objective is to educate and train. This particular video genre is important for e-learning and is often a simpler category of video as the aim of the video is unambiguous, and the techniques for educating and training fall in a somewhat well contained set, as opposed to a much larger set of techniques for handling aesthetics in motion pictures. The narrative in such video is also simpler, often linear, and mostly developed in a hierarchic way to elucidate concepts in greater levels of detail. Deductive approaches will move from the general to the particular, whilst inductive approaches will move from the detail to the general.The set of conventions and rules that shape these videos is termed production grammar, and our earlier work [1] has explored the structural elements. In this work we present a coherent approach using the hierarchical HMM to extract the structural units that form the building blocks of an education/training video. Rather than using the hand-crafted approach of our previous work [1] to define the structural units, we use the data to learn the parameters of a HHMM, and thus naturally extract the hierarchy. We then study this hierarchy and examine the nature of the structural units at different levels of abstraction.

Discovering structure in videos is a current active research area. The central task is to efficiently index data into semantic units, possibly at different level of abstractions, to simplify the process of navigation. There have been several systems proposed for specific video genres. Partition and classification of broadcast videos into meaningful sections has attracted significant attention. For instance, in [2], low-level features are combined with the concept of shot syntax to identify and label different narrative structures such as anchor shots, voice-over segments and interview sections found in news programs. Research into the domain of lecture videos be found in [3]. In their work, visual events are detected from the visual stream and then incorporated with audio information in a probabilistic framework to detect topic transitions. The domain of entertainment film has also been studied lately. Wang *et al.* [4], for example, attempt to detect *scenes* in film using the similarity in visual information and further improve the results with guidance from cinematic grammar.

In all videos there is often a common and natural hierarchy in the content. For example, surveillance videos map to different regions; news reports are organised into layers of details; a film has episodes, story units, scenes then shots, and a training video has several sub-topics which in turn form a main topic. Being able to model the hierarchic nature of a video offers ways to index the content in a meaningful hierarchy of semantic units. The hierarchical HMM has emerged naturally as a candidate to model this problem. The HHMM is a powerful stochastic model, first introduced in [5], in which the HHMM is viewed as a form of probabilistic context free grammar (PCFG), and the inference algorithm and parameter learning procedures are constructed based on the inside-outside algorithm. In [6], the HHMM is converted to a DBN, and a general DBN inference algorithm is applied to achieve complexity linear in time T, but exponential in the depth D of the HHMM. The same approach is applied in [7], ie: the HHMM is 'flattened' into regular HMM with a very large state space for inference purpose. Their work [7] aims to detect structures of soccer videos in an unsupervised manner. The model selection is first carried out using the MCMC to determine the structure parameters for the model, followed by a feature selection procedure. Finally, the HHMM is used to detect two semantic concepts, namely *play* and *break* in soccer videos. As there is little hierarchy at this level, the power of the hierarchic probabilistic model is not used. The HHMM is also applied in other domains other than multimedia such as in hand-written recognition [5], robot navigation, behaviour recognition and information retrieval.

The remainder of the paper is organised as follows. We briefly

0-7803-8554-3/04/$20.00 ©2004 IEEE.

discuss the HHMM with shared structures in Sec. 2. Next, Sec. 3 presents a brief analysis of the narrative structure in educational videos. This is followed by the experimental results in Sec. 4. Finally, the conclusion is provided in Sec. 5

2. THE HIERARCHICAL HMM WITH SHARED STRUCTURES

The discrete HHMM was originally proposed in [5] and its extension to accommodate shared structured has been addressed in our previous work [8]. We refer readers to [8] for a complete treatment. Here, we refocus our attention to elucidate the idea of the shared structures and briefly discuss the DBN representation and the EM procedure for parameter estimation. We then present the case when the emission probability is modeled as a mixture of Gaussians.

A hierarchical HMM extends the traditional HMM to allow each state itself to be a HMM. Formally, a HHMM is defined by a topological structure Γ and a set of parameter θ attached to the topology. The depth D –defines the number of layers in the hierarchy– and the *number of states* available at each level Q^d, for $d = 1, .., D$, are specified by Γ. Fig. 2 shows an example of a topology of depth 3. Level 1 is the root level and is always fixed to have only a single state, and only the lowest level D, termed as *production level*, emits observation. Furthermore, the topological structure reveals the 'parent-children' relationship of states between two consecutive levels[1]. A state p at level d is assigned to a *set of children*, $\text{ch}(p)$, at level $d+1$. A state i at level $d+1$ therefore might multiply inherit, or be *shared*, from more than one parent at level d. The general form of DBN representation[2] is shown in Fig. 1(a).

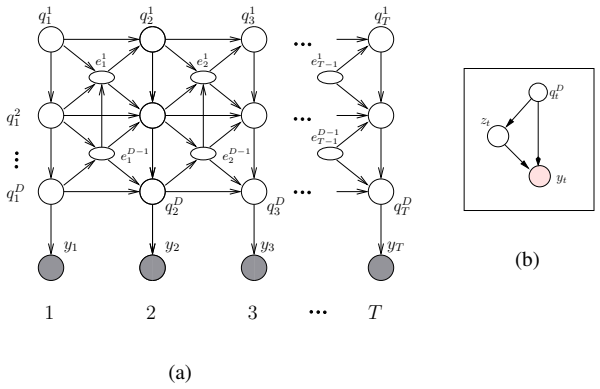

Fig. 1. (a) DBN representation for the discrete HHMM; (b) Added mixture component z_t at level D.

Given a topology Γ, the parameter $\theta \triangleq \{\pi, \text{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ of the HHMM is specified as follows.. For each level d (except D), $p \in Q^d$, and $i, j \in \text{ch}(p)$ we define: $\pi_i^{d,p}$ is the initial probability of the child i given the parent is p at level d; $\text{A}_{i,j}^{d,p}$ is the transition probability from child i; and $\text{A}_{i,\text{end}}^{d,p}$ is the probability that state p terminates at level d given its current child is i. We require that $\sum_i \pi_i^{d,p} = 1$, $\sum_j \text{A}_{i,j}^{d,p} = 1$, and $\text{A}_{i,\text{end}}^{d,p} \leq 1$, which later become Lagrangian constraints used in the maximisation step in the EM procedure. Finally, at level D an observation probability matrix B is specified in the case of discrete HHMM, or a set of $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ in case the observation values are continuous and modeled as a Gaussian mixture model (GMM).

Given an observed data set $\mathcal{D}$ and some initial parameters, the EM algorithm iteratively re-estimates a new parameter $\hat{\theta}$, hill climbing in the parameter space which is guaranteed to converge to a local maxima. As shown in [8], doing EM parameter re-estimation reduces to first calculating the expected sufficient statistics (ESS) $\bar{\tau} = \text{E}_{\mathcal{V}\setminus\mathcal{D}}\, \tau$, and then setting the re-estimated parameter $\hat{\theta}$ to the normalized value of $\bar{\tau}$. The ESS for parameter $\left\{\text{A}_{i,j}^{d,p}\right\}$, for example, is calculated as:

$$\bar{\tau}(\text{A})_{i,j}^{d,p} = \mathop{\text{E}}_{\mathcal{V}\setminus\mathcal{D}} \tau(\text{A})_{i,j}^{d,p} = \sum_{t=1}^{T-1} \xi_t^{d,p}(i,j) \Big/ \Pr(\mathcal{D}) \qquad (1)$$

where the auxiliary variable $\xi_t^{d,p}(i,j)$ is defined as the transition probability $\Pr(q_{t+1}^{d+1} = j, q_t^{d+1} = i, q_{t+1}^d = p, e_t^{d:d+1} = 01, \mathcal{D})$, which is diagrammatically visualised as $\left[{}_{i}\overset{p}{\Theta}{}_{j}\right]$. We refer readers to [8] for complete details on the computation of auxiliary variables and other expected sufficient statistics. The rest of this section will discuss the case when the emission probability is modeled as a GMM.

The general method when modeling GMM as the observation probability in the hierarchical HMM is similar to that of the regular HMM. The DBN structure at level D is modified as in Fig 1(b), where a mixture variable z_t is added. For simplicity, thereafter in this section we will drop the index D. Let M be the number of mixtures and N be the number of states at level D. The observation matrix B in the discrete case is replaced by the mixing weight matrix $\{\lambda_{mi}\}$ and a set of means and covariance matrices $\{\mu_{mi}, \Sigma_{mi}\}$ for $i = 1, \ldots, N$ and $m = 1, \ldots, M$. Given observed data $\mathcal{D}$, following a standard EM procedure, the expected complete log-likelihood $\langle \ell(\theta; \mathcal{D}) \rangle$ in the E-step is given as (discarding terms irrelevant to z_t and y_t):

$$\sum_{\substack{1 \leq i \leq N \\ 1 \leq m \leq M}} \left[\sum_{t=1}^{T} \left\langle \text{I}_{m,i}^{z_t, q_t} \right\rangle \log \mathcal{N}(y_t \mid \mu_{mi}, \Sigma_{mi}) + \sum_{t=1}^{T} \left\langle \text{I}_{m,i}^{z_t, q_t} \right\rangle \log \lambda_{mi} \right]$$

where $\mathcal{N}(.)$ is the multivariate Gaussian density function, $\langle . \rangle$ is the expectation operator, and $\text{I}_{m,i}^{x,y}$ is the identity function which takes 1 if $\{x = m \cup y = i\}$, and 0 otherwise. It is straightforward to compute the expectation of the identity function:

$$\begin{aligned} \left\langle \text{I}_{m,i}^{z_t, q_t} \right\rangle &= \Pr(z_t{=}\, m, q_t{=}\, i \mid \mathcal{D}) = \Pr(z_t{=}\, m \mid q_t{=}\, i, y_t) \Pr(q_t{=}\, i \mid \mathcal{D}) \\ &= \frac{\lambda_{mi} \mathcal{N}(y_t \mid \mu_{mi}, \Sigma_{mi})}{\sum_{m=1}^{M} \lambda_{mi} \mathcal{N}(y_t \mid \mu_{mi}, \Sigma_{mi})} \times \frac{\gamma_t^D(i)}{\Pr(\mathcal{D})} \end{aligned}$$

where[3] the auxiliary variable $\gamma_t^D(i)$ defined as the probability $\Pr(q_t^D{=}\, i, \mathcal{D})$ and can be computed directly from horizontal transition probability $\xi_t^{d,p}(i,j)$ and vertical transition probability $\chi_t^{d,p}(i)$ (defined [8]). Finally, in M-step we maximise the expected complete log-likelihood $\langle \ell(\theta; \mathcal{D}) \rangle$ with respect to λ_{im} and μ_{mi}, Σ_{mi}. Introducing Lagrange multipliers for λ_{mi} and setting derivatives to zero for μ_{mi}, Σ_{mi}, the set of re-estimated parameters is given

[1] Note that the original HHMMin [5] assumes that a state has a only a single parent and therefore the topology reduces strictly to a tree.

[2] The DBN structure for HHMM was originally presented in [6]

[3] we put back hierarchic index D for clarity here

as:

$$\hat{\varepsilon}_{mi} = \frac{\sum_{t=1}^{T} \left\langle \mathrm{I}_{m,i}^{z_t,q_t} \right\rangle}{\sum_{m=1}^{M} \sum_{t=1}^{T} \left\langle \mathrm{I}_{m,i}^{z_t,q_t} \right\rangle}, \qquad \hat{\mu}_{mi} = \frac{\sum_{t=1}^{T} \left\langle \mathrm{I}_{m,i}^{z_t,q_t} \right\rangle y_t}{\sum_{t=1}^{T} \left\langle \mathrm{I}_{m,i}^{z_t,q_t} \right\rangle}$$

$$\hat{\Sigma}_{mi} = \frac{\sum_{t=1}^{T} \left\langle \mathrm{I}_{m,i}^{z_t,q_t} \right\rangle (y_t - \mu_{mi})(y_t - \mu_{mi})^{\mathsf{T}}}{\sum_{t=1}^{T} \left\langle \mathrm{I}_{m,i}^{z_t,q_t} \right\rangle}$$

When multiple (iid) observation sequences are given, the set of above equations can be adjusted by simply adding a summation over the number of sequences. This corresponds to 'counting' over all sequences.

3. NARRATIVE STRUCTURAL UNITS

The structure of an education/training video is fashioned by the way in which training material is presented. In almost all cases, there is a narrator that takes us on the journey of learning through the subject. Thus, the narrator will reappear through the video, guiding us along the way. To make the points, the narrator uses a bag of tricks: text, footage with voice-over, or discussion with people or interviews. Generally a shot in this video genre can be classified into one of following three categories. *On-screen* narration refers to sections of narration, both direct and assisted, in which the narrator directly speaks to the audience. *Voice-over* refers to sections where the narrator will maintain the narrative by using voice-over. *Linkage* sections refers to sections of informational footage without the narrator and we divide this into two subgroups: *expressive* linkage and *functional* linkage, where the former is used to dramatise a situation. Readers are referred to [1] for precise definitions and examples of these structural units. For now, it is sufficient for us to augment these four main categories as a set of high level semantics. In a generative process, one can imagine that they correspond to four parent states that consists of a number children states producing what we see in the video such as the red colour, the caption texts and so on. We use this knowledge to construct a topology for the experiment presented next.

4. EXPERIMENTAL RESULTS

In this experiment, we present initial results on applying the HHMM to automatically map semantic concepts to the model states at different levels of abstraction. Taking the advantages of the expressiveness of shared structures, we construct a 3-level HHMM with a 'true' model in mind, ie: the topology of the HHMM is constructed in such a way that it maps into the hypothesised hierarchy as in Fig 2. That is, for example, at the production level, we are hypothesise that the model can 'perfectly' learn to map each state to an elementary feature such as face, text or speech, from which higher levels of descriptions such as *on-screen* narration section can be built.

Given this topology, a 3-level HHMM is subsequently formed. The production level includes 7 states directly attached to the observed features (recall that these are shot-based features). The next level have 4 states, which in our opinion, adequately reflects the number of higher level units in this experiment. Finally, the root covers the entire video. We first randomly initialise the parameter of the model and then use EM to learn a new set of model parameters.

The feature vector used in this experiment includes seven features computed at the shot level. Three of them are from the visual stream, namely: face-content-ratio, text-content-ratio, and average motion based on camera pan and tilt estimation; and the other four features are extracted from the sound track including: music-ratio, speech-ratio, silence-ratio, and non-literal sound (NL-) ratio. Computation of these features are detailed in [9]. Note that all of these features are in the range of $[0, 1]$.

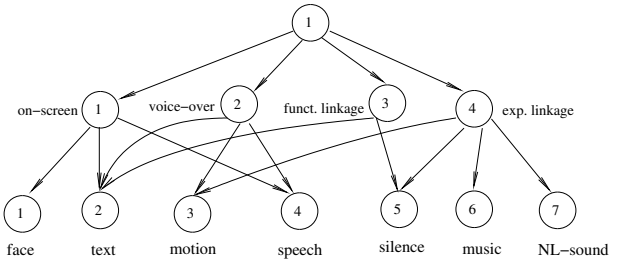

Fig. 2. Hierarchy of connected concepts which is used as topological specification for the corresponding HHMM (links from state $1, 2$ (level 2) to state 7 (level 3) are not shown for readability).

Let V be any arbitrary video, and T be the number of shots in V. Feature extraction from V will result in a sequence of observations of length T where each observation is a column vector of seven features. We collect nine videos in total to use for the training purpose with T ranging from 124 to 245. We analyse the learned model and present the results at two levels.

Semantic mapping at the production level
We use the mean matrix at the production level to understand the mapping between the production level states and the feature vector components. Figure 3 shows the visualisation of the means. Thus, for example, state 1 is strongly linked to face and speech, and a little bit of motion. This state corresponds to direct and assitive narration sections. State 2 corresponds to speech and motion, and thus to voice-over sections. State 3 corresponds to NL-sound, and thus to expressive linkage sections and so on. In Table 1, we summarise this result of mapping the production level states to the structural units manually crafted in our earlier work. The mean values learned at the production level is visualised in Fig 3.

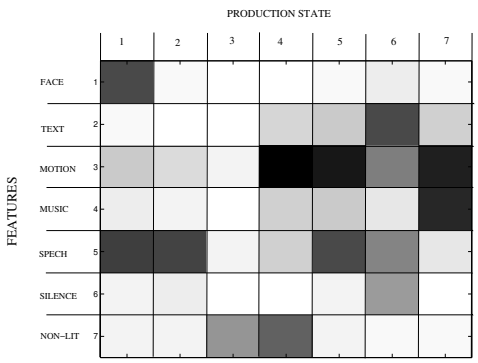

Fig. 3. Gray-scaled visualisation of the mean values of seven features learned at the production level.

Semantic mappings at the upper level
The middle level of topology (Fig. 2) contains 4 states. Combining the information of the 'parent-child' relationship specified in topology and results from Table 1, we can interpret the meaning of

State	Feature	Structural Units
1	face, speech, motion	direct-, assistive- narration
2	speech, motion	*pure* voice-over (VO)
3	NL-sound	expressive linkage
4	motion, NL-sound, text	linkage, VO with texts
5	motion, speech, music	expressive linkage
6	text, motion, speech	VO, functional linkage
7	motion, music	dramatic/expressive linkage

Table 1. Deduced semantic mappings at production (shot) level.

each state. However, this must be done in conjunction with the estimated π matrix since it specifies the probability of a child being called.

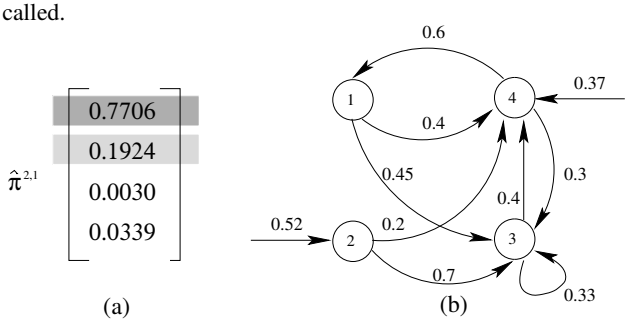

Fig. 4. Estimated transition probability at the upper level. Only dominant probabilities are shown.

For example, from Fig 2, state 1 has four children at the production level, namely $1, 2, 4$ and 7. However, the estimated initial probability $\hat{\pi}^{2,1}$ shown in Fig 4(a) indicates that the children 4 and 7 are almost disconnected.This licenses us to attach this parent state with only on-screen section or with pure voice-over section (with a lower probability). Similar analysis shows that state 2 is connected with expressive linkage sections or voice-over; state 3 is connected with expressive voice-over (ie: voice over with no texts and lots of motion); state 4 is connected with functional linkage or sometimes with voice-over sections. Overlapping is observed with the voice-over sections, however, they have different probabilities. A transition matrix for these four states is shown in Fig 4(b). While we cannot deduce the complete 'style' of educational video from this result, it does suggest a 'common' presentational style for this genre of video. In this case, written in regular grammar, it is[4]:

$$[2|4][4,1]^*[3]^+[4,1]^*[1,4]^*[3]^+$$

At this stage, we cannot make conclusions about the power of this type of grammar[5]. However, our initial results from the Viterbi decoding suggests a very interesting point. The sequence of states and time indices decoded at the middle level for the video "Eye-Safety" is: $\begin{bmatrix}\mathbf{4}\\1\end{bmatrix}\begin{bmatrix}\mathbf{1}\\5\end{bmatrix}\begin{bmatrix}\mathbf{4}\\14\end{bmatrix}\begin{bmatrix}\mathbf{1}\\15\end{bmatrix}\begin{bmatrix}\mathbf{4}\\16\end{bmatrix}\begin{bmatrix}\mathbf{3}\\17\end{bmatrix}\begin{bmatrix}\mathbf{4}\\18\end{bmatrix}\begin{bmatrix}\mathbf{1}\\22\end{bmatrix}\begin{bmatrix}\mathbf{4}\\23\end{bmatrix}\begin{bmatrix}\mathbf{1}\\24\end{bmatrix}$ $\begin{bmatrix}\mathbf{4}\\25\end{bmatrix}\begin{bmatrix}\mathbf{3}\\28\end{bmatrix}\begin{bmatrix}\mathbf{4}\\44\end{bmatrix}\begin{bmatrix}\mathbf{1}\\45\end{bmatrix}\begin{bmatrix}\mathbf{4}\\49\end{bmatrix}\begin{bmatrix}\mathbf{1}\\50\end{bmatrix}\begin{bmatrix}\mathbf{3}\\52\end{bmatrix}\begin{bmatrix}\mathbf{4}\\57\end{bmatrix}\begin{bmatrix}\mathbf{1}\\58\end{bmatrix}\begin{bmatrix}\mathbf{4}\\65\end{bmatrix}\begin{bmatrix}\mathbf{1}\\67\end{bmatrix}\begin{bmatrix}\mathbf{4}\\70\end{bmatrix}$ $\begin{bmatrix}\mathbf{3}\\71\end{bmatrix}\begin{bmatrix}\mathbf{4}\\72\end{bmatrix}\begin{bmatrix}\mathbf{3}\\73\end{bmatrix}\begin{bmatrix}\mathbf{1}\\91\end{bmatrix}\begin{bmatrix}\mathbf{3}\\94\end{bmatrix}\begin{bmatrix}\mathbf{4}\\103\end{bmatrix}\begin{bmatrix}\mathbf{1}\\104\end{bmatrix}\begin{bmatrix}\mathbf{3}\\109\end{bmatrix}\begin{bmatrix}\mathbf{4}\\117\end{bmatrix}\begin{bmatrix}\mathbf{1}\\118\end{bmatrix}\begin{bmatrix}\mathbf{3}\\119\end{bmatrix}$ $\begin{bmatrix}\mathbf{1}\\124\end{bmatrix}$, where each entry is a 2-tuple $\begin{bmatrix}\mathbf{s}\\t\end{bmatrix}$ where s is the state activated at time t. A useful state pattern from the above the grammar is $[41]$ or $[14]$. We observe this pattern in this video at times $16, 23, 25, 45, \ldots$ or 118. These correspond to sub-topic boundaries. In fact, when compared with the ground-truth there is only one miss and one false positive.

[4]Note that together with the probabilities specified in Fig 4(b), this grammar should be viewed as Probabilistic Regular Grammar.

[5]We also notice the absence of state 2 in the middle part of this grammar. This is somewhat due to the rare appearance of expressive linkage sections except at the starting part used to build an introduction 'mood'

5. CONCLUSION

In this paper we have utilised the hierarchic expressiveness of the hierarchical HMM to automatically exploit the structural units in an education and training video. We extend the HHMM to deal with continuous observations, and estimate the parameters of a 3-level HHMM for this task using nine training videos. We have analysed the hierarchy constructed in this data driven approach, and our results indicate the promise of this approach for structure discovery in video data.

6. REFERENCES

[1] Dinh Quoc Phung, Chitra Dorai, and Svetha Venkatesh, "Narrative structure analysis with education and training videos for e-learning," in *International Conference on Pattern Recognition*, Quebec, Canada, 11-15 August 2002, pp. 835–839.

[2] K. Shearer, C. Dorai, and S. Venkatesh, "Incorporating domain knowlege with video and voice data analysis," 2000, MDM/KDD 2000, Workshop on Multimedia Data Minning, Aug 20-23, Boston, USA.

[3] Tanveer Seyeda-Mahmood and S. Srinivasan, "Detecting topical events in digital video," in *ACM Multimedia*, 2000, pp. 85–94.

[4] Jihua Wang, Tat-Seng Chua, and Liping Chen, "Cinematic-based model for scene boundary detection," in *The Eighth Conference on Multimedia Modeling*, Amsterdam, Netherland, 5 - 7 Nov. 2001.

[5] Shai Fine, Yoran Singer, and Nftali Tishby, "The hierarchical hidden markov model: Analysis and applications," *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.

[6] Kelvin Murphy and M. Paskin, "Linear time inference in hierarhical hidden markov models," in *Proceedings of Neural Information Processing Systems*, Vancouver, Canada, 2001.

[7] Lexing Xie, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun, "Unsupervised discovery of multilevel statistical video structures using hierarhical hidden markov models," in *IEEE International on Multimedia and Expo*, Baltimore, USA, July 6-9 2003, pp. III.29 – III.32.

[8] Hung H. Bui, Dinh Q. Phung, and S. Venkatesh, "Hierarchical hidden markov models with general state hierarchy," in *submitted to The Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, San Jose, California USA, July 25-29 2004.

[9] Dinh Quoc Phung, Svetha Venkatesh, and Chitra Dorai, "On extraction of thematic and dramatic functions in educational films," in *IEEE International Conference on Multimedia and Expo*, Baltimore, New York, USA, 6-9 July 2003, pp. 449 – 452.