LAYERED DYNAMIC MIXTURE MODEL FOR PATTERN DISCOVERY IN ASYNCHRONOUS MULTI-MODAL STREAMS

Lexing Xie[†], *Lyndon Kennedy*[†], *Shih-Fu Chang*[†], *Ajay Divakaran*[§], *Huifang Sun*[§], *Ching-Yung Lin*[‡]

[†]Dept. of Electrical Engineering, Columbia University, New York, NY
[§]Mitsubishi Electric Research Labs, Cambridge, MA
[‡]IBM T. J. Watson Research Center, Hawthone, NY

ABSTRACT

We propose a layered dynamic mixture model for asynchronous multi-modal fusion for unsupervised pattern discovery in video. The lower layer of the model uses generative temporal structures such as a hierarchical hidden Markov model to convert the audio-visual streams into mid-level labels, it also models the correlations in text with probabilistic latent semantic analysis. The upper layer fuses the statistical evidence across diverse modalities with a flexible meta-mixture model that assumes loose temporal correspondence. Evaluation on a large news database shows that multi-modal clusters have better correspondence to news topics than audio-visual clusters alone; novel analysis techniques suggest that meaningful clusters occur when the prediction of salient features by the model concurs with those shown in the story clusters.

1. INTRODUCTION

This paper is concerned with the discovery of meaningful patterns across multiple modalities in video. Patterns refer to the recurrent segments with consistent characteristics in temporal sequences. Unsupervised discovery aims to recover a statistical description of the structures and segment the data without prior labeling and training, this is preferred when the meanings are copious yet unknown a priori. Discovering complex patterns in video is particularly challenging since they are often defined on several input modalities, across different levels of semantics, and over a period of time. Inspecting a news collection on "1998 NBA Finals" for example, we will find basketball highlights with fast dynamics, stadium and interview segments, along with words and phrases identifying the time and the stage of the games.

Prior work addressed unsupervised pattern discovery on one input stream, such as finding the latent dimensions in text [4] or mining temporal patterns from audio-visual observations [10]; multi-modal fusion also appeared in various contexts, such as audio-visual speech recognition [1] where the audio and video are in exact sync, and cross-media annotation [3] where different modalities bear the same set of meanings. None of these models readily handles multi-modal fusion across the asynchronous and semantically diverse audio, visual and text streams, hence we propose a layered dynamic mixture model to this end. The model first groups each stream into a sequence of mid-level labels so as to account for the temporal dependency and noise; it then infers a high-level label from the mid-level labels in all streams, and addresses the cross-stream asynchrony by allowing loose correspondence across different modalities. This layered structure is similar to the layered HMM for multi-modal user interaction [2], except that the layered HMM only handles fixed-rate inputs and enforces rigid correspondences for different modalities. Our use of text information resembles those in multimedia annotation [3, 11], except the correlations in text are explicitly modeled, and the text information directly influences the composition of the multi-modal clusters as opposed to just serving as the explanation.

We evaluate this model on TRECVID 2003 broadcast news videos. Results show that the multi-modal clusters have better correspondence to news topics than audio-visual clusters; on a subset of topics that bear salient perceptual cues, they have even better correspondence than text. Manual inspection of the multi-modal clusters finds a few consistent clusters that capture the salient syntactic units in news broadcast, such as the financial, sports or commercial sections. Furthermore, these clusters tend to appear when the mixture model is able to predict audio-visual features and words that are indeed salient in the actual story clusters.

In Section 2 we present the formulation and inference algorithm for the the layered dynamic mixture model; in Section 3 we discuss the features used and the low-level clustering algorithms that correspond to observations in each modality; in Section 4 we show experiments on news videos; Section 5 presents the conclusion and a few extensions.

2. LAYERED DYNAMIC MIXTURE MODEL

The structure of the layered mixture model is shown in Fig. 1(a). Multi-modal fusion for unsupervised learning differs from those for supervised learning [8] in that neither labeled ground-truth nor class separability is available as the computational criteria for guiding the fusion model. Therefore we use the data likelihood in generative models as an alternative criterion to optimize the multi-level dynamic mixture model.

2.1. The layered representation

The layered dynamic mixture representation consists of the low-level feature streams, the mid-level labels and the high-level fused clusters, and the two layers of probabilistic models in between. Aside from enhancing the robustness and reducing parameter tuning [2], introducing layers in unsupervised learning has intuitive and mathematical advantages. The layers divide the problem into modeling noise and temporal dependencies in the individual modalities and fusing the modalities of different temporal resolution. This separation enables the use of different model structures in the lower layer so as to take advantage of the domain knowledge in each individual modality. In terms of representation, a layered model is more flexible and yields better clustering results than one-level clustering as seen in Section 4.2.

The layered mixture model has a set of different temporal indexes. We can see from Fig. 1(a) that the layered mixture model

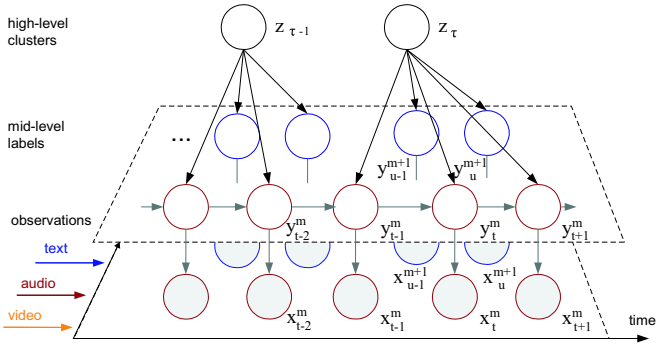


Fig. 1. The layered dynamic mixture model, in the dimensions of time, various modalities and perceptual levels. Shaded nodes: observed, clear nodes: hidden; different colors represent nodes in different modality.

allows different temporal index t^m for each input modality m. The lower layer models group each input stream x_t^m into a mid-level label stream y_t^m using generative models tailored to each modality (Section 3.2), and y_t^m has the same temporal resolution as x_t^m. We further partition the time axis into a set of non-overlapping loose temporal *bags* τ, and this bag contains a number of continuously indexed samples in each stream (assume non-empty without loss of generality), denote as $\{t^m \mid t^m \in \tau\}$. We assign one top layer node z to each bag, and with a slight abuse of notation, we also use τ to index the top-layer nodes, written as z_τ.

2.2. The fusion layer

In this work, the temporal bag boundaries lie on syntactically meaningful boundaries in the video, such as scene transition points or news stories boundaries. For efficiency, we perform clustering in each layer separately, i.e., the values of mid-level labels y are taken as the maximum-likelihood value from the lower-layer model and considered as "observed" when inferring the high-level node z.

Denote the set of mid-level labels within each bag as $y_\tau = \cup_m \{y_t^m, t^m \in \tau\}$. For simplicity, we assume conditional independence among the labels in y_τ given z_τ across both the time and the modalities, i.e.,

$$p(y_\tau, z_\tau) = p(z_\tau) \prod_m \prod_{t^m \in \tau} p(y_t^m | z_\tau) \quad (1)$$

Under this assumption, the temporal orders within each bag no longer influence the value of z_τ; when each of y_t^m takes discrete values, y_τ can be represented by a set of multi-dimensional co-occurrence counts $c(m, \tau, y) = \#\{y_t^m = y, t^m \in \tau\}$. Intuitively, this is to treat the mid-level labels y^m, obtained by *de*-noising and de-correlating x^m, as if they were generated by independent multinomial draws conditioned on the high-level *meaning* z.

According to this definition, we rewrite the complete-data log-likelihood of y and z in bag τ from Eq. (1), and estimate the model parameters with the EM algorithm:

$$\begin{aligned} l(\tau, z) &\triangleq \log p(y_\tau, z_\tau = z) \\ &= \log p(z) + \sum_{(m,y)} c(m, \tau, y) \log p(y^m | z). \end{aligned} \quad (2)$$

The E-step reads:

$$p(z_\tau = z | y_\tau) = \frac{\exp l(\tau, z)}{\sum_{z \in \mathcal{Z}} \exp l(\tau, z)} \quad (3)$$

The M-step follows:

$$p^*(z) = \frac{1}{T} \sum_{\tau=1}^{T} p(z_\tau = z | y_\tau) \quad (4)$$

$$p^*(y^m | z) = \frac{\sum_{\tau=1}^{T} c(m, \tau, y) p(z_\tau = z | y_\tau)}{\sum_y \sum_{\tau=1}^{T} c(m, \tau, y) p(z_\tau = z | y_\tau)} \quad (5)$$

We can extend this basic fusion model to include a joint inference from observations x_t^m to the highest level z_τ, to model dependency within each temporal window, or to allow flexible temporal bags to be learned while performing model inference.

3. PROCESSING MULTI-MODAL INPUT

We now ground the layered dynamic mixture model on video data. We describe the extraction of multi-modal feature streams and the choices of generative models used for mid-level grouping. We use news videos as our test domain, while the discussions can be generalized to other domains with salient syntactic structures.

News videos are semantically rich and syntactically constrained. The basic temporal units in news videos are *shots* and *stories*. A *shot* refers to a continuous camera take in both time and space. A *story* is defined [6] as a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses. Shot boundaries in news can be reliably detected with over 90% accuracy, while state-of-the-art audio-visual story segmentation has an F1 measure $\sim 75\%$ [8].

3.1. Multi-modal features

We extract from each video the following sets of low-level audio-visual descriptors, visual concepts, and text terms as the base layer in the hierarchical mixture model:

(1) Color Histogram of an entire frame is obtained by quantizing the HSV color space into fifteen bins: white, black and gray by taking the extreme areas in brightness and saturation; equal-width overlapping bins on the hue values resembling the six major colors red, yellow, green, cyan, blue and magenta in high and low saturations, respectively. Yellow with low saturation is then excluded to make a linearly independent 14-dimensional feature vector averaged over a time window of one second.

(2) Motion intensity consists of the average of motion vector magnitude and the least-square estimate of horizontal pan from the MPEG motion vectors, extracted every second.

(3) The audio features contain a four-dimensional vector every half a second: the mean *pitch* value; the presence/absence of *silence* and *significant pause*, the latter obtained by thresholding locally normalized pause length and pitch jump values; six-class *audio category* labels from a GMM-based classifier (silence, female, male, music, music+speech, or other noise) [9].

(4) A visual concept vector contains the confidence scores for how likely the keyframe of a shot is to contain a set of visual concepts [7]. The concepts used in this work are pruned from a lexicon of over a hundred requiring them being as specific as possible while having reasonable detection accuracy. They include: five events - *people, sport, weather, cartoon, physical violence*; five scenes - *studio setting, non-studio setting, nature-vegetation, nature non-vegetation, man-made scene*; twelve objects - *animal, face, male news person, male news subject, female news person, female news subject, person, people, crowd, vehicle, text overlay, graphics*.

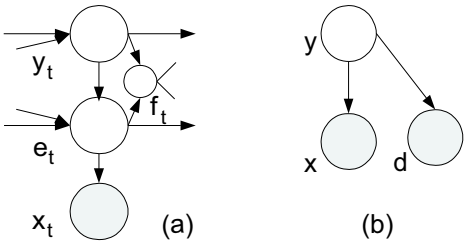


Fig. 2. Models for mapping observations to mid-level labels. (a) Hierarchical HMM; (b) PLSA.

(5) Keyword features can be obtained from either the closed captions or automatic speech recognition (ASR) transcripts. Stemming, part-of-speech tagging and rare word pruning are performed, retaining a 502-token lexicon of frequent nouns, verbs, adjectives and adverbs from the TRECVID corpus. The tf-idf score of the words within a news story are used as the feature vector.

3.2. Unsupervised mid-level grouping

The audio-visual features (1)-(3) are sampled over uniform intervals, (4) over each shot, and the text features are taken for every story. The audio-visual streams exhibit temporal dependency [10], while independence between the keywords in adjacent stories can reasonably be assumed as they are often on different topics. For the former we use hierarchical hidden Markov model (HHMM) for unsupervised temporal grouping, as described in our earlier work [10, 11]; for the latter we use probabilistic latent semantic analysis (PLSA) [4] to uncover the latent semantic aspects.

The unsupervised temporal grouping algorithm [10] takes a collection of unlabeled feature streams, learns an HHMM (Fig. 2(a)), and at the same time associates the most likely label sequences with the streams. The size of the model and an optimal feature set are automatically selected using the mutual information and Bayesian information criteria (BIC), and the resulting HHMM typically has two to four top-level states. To incorporate news story boundaries, the HHMM inference algorithm only allows highest-level state transitions at story boundaries, hence restricting the segment coming from the same story to stay in the same branch of the Markov chain hierarchy.

The PLSA model is shown in Fig. 2(b). Observing the words x grouped in stories d, the model learns the conditional dependencies between the hidden semantic aspects y and both observed variables. The double mixture structure of PLSA provides more flexibility for capturing word distributions than a simple mixture model, we have observed that the PLSA distributions more accurately capture sets of semantically related words, rather than being deluged by the frequent words.

4. EXPERIMENTS

We test the proposed fusion model on TRECVID news corpus [5]. This data set contains 151 half-hour broadcasts of *CNN Headline News* and *ABC World News Tonight* from January to June, 1998. The videos are encoded in MPEG-1 with CIF resolution; also available are the ASR transcripts produced by LIMSI. We partition the dataset into four quadrants each containing about half of the videos from one channel.

After feature extraction, one HHMM is learned on each of the color, motion and audio features; the visual concepts, due to their diverse nature, yield three HHMMs grouped by the automatic feature selection algorithm; the words in all stories are clustered into 32 latent dimensions with PLSA. The fusion model then infers a most-likely high-level hidden state from all the mid-level states in each story, taking one of 32 possible values. The multi-level clustering algorithm runs in linear-time, and it typically takes less than three hours on a 2GHz PC for 19 hours of video.

4.1. Inspecting the clusters

We first inspect a story-board layout of each cluster. As seen from Fig. 3(a), the weather forecast cluster results from the consistent color layout and similar keywords, while the sports cluster is characterized by the motion dynamics and visual concepts regarding people or person. Other interesting clusters include: a concentration of CNN financial news (precision 13/18 stories) with similar graphics, anchor person and common transitions between them; seven commercial sections in an eight-story cluster characterized by audio-visual cues only. These observations suggest that the multi-modal clusters are formed because of the consistency in appearance and syntax at the level of general content categories.

We also inspect the distribution of the multi-modal features. We compare the most likely feature distribution predicted by the model and those observed in the actual story clusters. An agreement between these two on one feature suggests that this may be a salient dimension that the model manages to capture. In Fig. 3(b) we plot the predicted and observed cluster-feature pairs with high statistical confidence into two color channels by applying simple cut-off rules for having high confidence. We require: P$>$0.8 for both the top-level mixture model $p(y^m|z)$ and the emission probabilities $p(x|y)$ of the discrete-valued features, a peak greater than twice the second-highest mode for the continuous features described by a mixture of Gaussians, or the intersection of the top 20 words in the PLSA probabilities $p(x|y)$ with those in the text transcript. We can see that the sports cluster in Fig. 3(a) shows high confidence in both the model and the observations for the visual concepts *animal* and *male-news-subject*; while the two weather clusters predict larger amounts of yellow and blue in the color layout and a set of weather-related words. We can also interpret from the common keywords that some clusters are a mixture of two or more general categories, e.g., politics and weather in cluster #27.

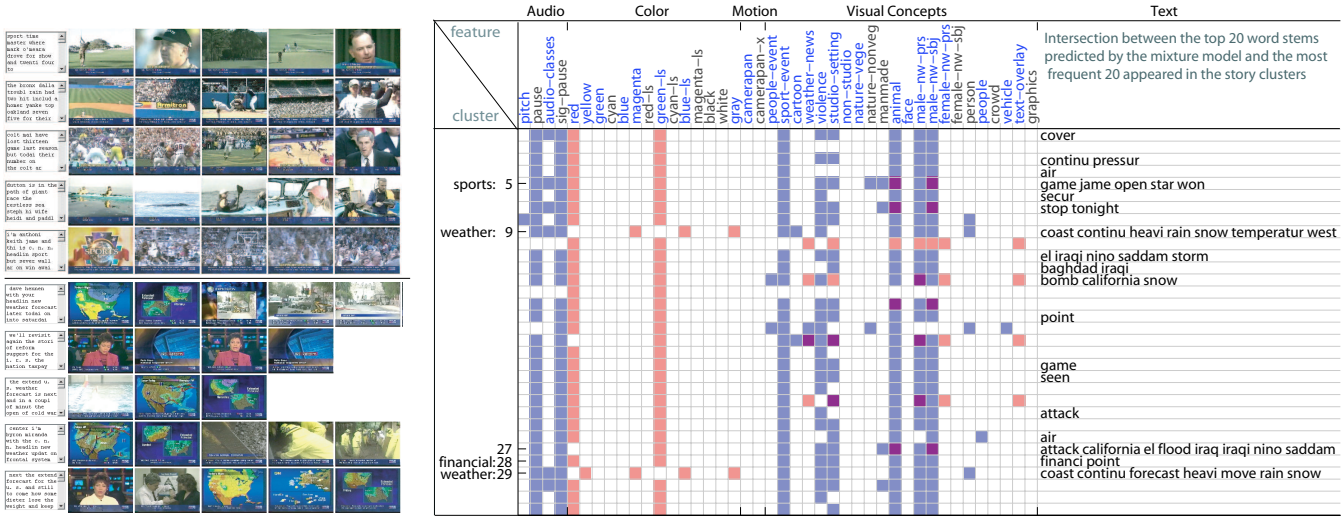
4.2. News topics

Topics are the semantic threads in news, defined as "an event or activity, along with all directly related events and activities" [6]. We compare the multi-modal clusters with the 30 topics present on the subset of labeled stories covering $\sim 15\%$ of the corpus. We use the TDT evaluation metric *detection cost*, a weighted sum of the precision and recall for each topic s:

$$C_{det}(s) = min_z\{P_M(z,s) \cdot P(s) + P_{FA}(z,s)(1 - P(s))\} \quad (6)$$

Here P_M and P_{FA} are the miss and false alarm probabilities of cluster z with topic s, i.e., $P_M = |s \cap \bar{z}|/|s|$ and $P_{FA} = |\bar{s} \cap z|/|\bar{s}|$; and $P(s)$ is the prior probability of topic s. The detection cost is then divided by $min\{p(s), 1 - p(s)\}$ to normalize against trivial decisions. Note the smaller the value of $\bar{C}_{det}$, the better a topic can be represented by one of the clusters.

We compare the layered fusion structure with a one-level clustering algorithm, the latter is obtained by substituting the observations x into the places of y in Eq.(2-5). The average detection cost for the former is lower than the latter by 0.240. Furthermore, the average detection cost of any single modal audio-visual clusters is higher than the multi-modal clusters by 0.227, where the former



(a) Example clusters, each row contains the text transcript and the keyframes from a story. Top: #5, sports, precision 16/24; Bottom: #29, weather, precision13/16.

(b) The most probable features predicted by the model (red) and observed in the clusters (blue). A magenta cell results from the joint presence from both, and a blank one indicates that neither has high confidence. The features retained during the automatic feature selection process are shown in navy.

Fig. 3. Observations in the multi-modal clusters. The models are learned on CNN set A, evaluated on set B.

is taken as the best corresponding model among all audio-visual HHMMs in Section 3.2, and equivalent to those in [11].

In seven out of the thirty topics, multi-modal clusters have lower detection costs than using text alone (i.e., the PLSA clusters), these topics include *1998 Winter Olympics*, *NBA finals* and *tornado in Florida* among others. The topics of improved performance tend to have rich non-textual cues, for instance the correspondence of cluster #5 in Fig.3(b) to *1998 Winter Olympics* can be explained by the prevalence of *music-speech* in the audio and *male-news-subject* in the visual concepts. Note the topics being evaluated are defined and labeled solely with the news transcripts [6], and this may implicitly bias the evaluation towards the meanings in text. For example, cluster #28 in Fig.3(b) corresponds to the topic *Asian economy crisis* while a better correspondence can be obtained with a PLSA cluster containing keywords *dow, nasdaq, dollar*, etc. Comparing the topic correspondence and the cluster inspection in Section 4.1 suggest that the multimodal clusters seem to be at a different level than the news topics, therefore the definition of a new set of ground truth from multimedia is called for.

5. CONCLUSION

We present dynamic layered mixture model for multi-modal fusion across asynchronous streams, this model uses a layered structure that can accommodate the diversity in the input streams and efficiently integrate the statistical evidence from all of them. Meaningful syntactic clusters are obtained on a large news video corpus using audio, visual and text features. Future work include further analysis and interpretation of the model, improving the features, and finding novel applications for multi-modal analysis.

ACKNOWLEDGEMENT

We thank R. Radhakrishnan for the audio classfier, W. Hsu for the story segmentation results, the TRECVID team [7] at IBM T. J. Watson Research Center for the concept detection results, and M. Naphade and A. Natsev for providing additional visual concept detection results.

6. REFERENCES

[1] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP, Journal of Applied Signal Processing*, vol. 2002, p. 1274, November 2002.

[2] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for learning and inferring office activity from multiple sensory channels," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI'02)*, (Pittsburgh, PA), October 2002.

[3] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. SIGIR-03*, (Toronto, Canada), pp. 119–126, July 28– Aug. –1 2003.

[4] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. ACM SIGIR*, pp. 50–57, ACM Press, 1999.

[5] NIST, "TREC video retrieval evaluation (TRECVID)," 2001–2004. http://www-nlpir.nist.gov/projects/trecvid/.

[6] NIST, "Topic detection and tracking (TDT)," 1998–2004. http://www.nist.gov/speech/tests/tdt/.

[7] A. Amir, G. Iyengar, C.-Y. Lin, C. Dorai, M. Naphade, A. Natsev, C. Neti, H. Nock, I. Sachdev, J. Smith, Y. Wu, B. Tseng, and D. Zhang, "The IBM semantic concept detection framework," in *TRECVID Workshop*, 2003.

[8] W. H.-M. Hsu, S.-F. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *SPIE Electronic Imaging*, January 2004.

[9] R. Radhakrishan, Z. Xiong, A. Divakaran, and Y. Ishikawa, "Generation of sports highlights using a combination of supervised and unsupervised learning in audio domain," in *Proc. Pacific Rim Conference on Multimedia*, 2003.

[10] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, *Unsupervised Mining of Statistical Temporal Structures in Video*, ch. 10. Kluwer Academic Publishers, 2003.

[11] L. Xie, L. Kennedy, S.-F. Chang, A. Divakaran, H. Sun, and C.-Y. Lin, "Discover meaningful multimedia patterns with audio-visual concepts and associated text," in *Int. Conf. Image Processing (ICIP)*, October 2004.