# PROBABILISTIC GRAPHICAL MODEL FOR AUTO-ANNOTATION, CONTENT-BASED RETRIEVAL, AND CLASSIFICATION OF TV CLIPS CONTAINING AUDIO, VIDEO, AND TEXT

*D. Putthividhya*[1], *H. T. Attias*[2], *S. S. Nagarajan*[3], *T.-W. Lee*[1]

[1]Institute for Neural Computation, UCSD
9500 Gilman Drive,
La Jolla, CA 92093

[2]Golden Metallic, Inc.
P.O. Box 475608
San Francisco, CA 94147

[3]Dept. of Radiology, UCSF
513 Parnassus Ave,
San Francisco, CA 94143

## ABSTRACT

We present a probabilistic graphical model that learns the joint statistical structures of text, audio, and video for the purpose of classification and retrieval of multimedia documents. The proposed model, which we call Multi-modal LDA (MM-LDA), builds on the basic Latent Dirichlet Allocation (LDA) model by postulating common hidden factors, termed topics, that are shared among the 3 data modalities. These hidden topics correspond to patterns of word co-occurrences in multimedia documents and describe how text words co-occur with certain visual and acoustic features. We demonstrate the power of MM-LDA in representing TV clips containing closed-captions, video, and audio, and show promising results in 3 challenging applications: TV clip classification, retrieval, and auto-annotation.

***Index Terms***— Multi-modal data representation, Content-based video retrieval, Automatic video annotation, Multimedia processing, Multimedia information retrieval.

## 1. INTRODUCTION

Today's advanced digital media technology has led to the explosive growth of multimedia data in a scale that has never occurred before. The availability of such large-scale quantities of multimedia documents prompts the need for efficient algorithms to search and index multimedia files. Modern multimedia content is often characterized by having multiple varied forms, i.e. movies consisting of video-audio streams with text captions; web pages containing pictures, text, and songs. This heterogeneous multi-modal nature gives rise to challenging new research questions of how to best represent, classify, and effectively retrieve multimedia data. The tremendous potential of such aforementioned research in a wide array of applications has drawn considerable attention to the emerging field of multimedia information retrieval in recent years.

Most previous work on multi-modal data retrieval and classification, e.g. [1], assumes simplistically that different modalities of data are independent. Retrieval/classification is thus performed on each modality separately and the results are subsequently combined. Often, knowledge about one modality conveys a great deal of information about the others. Making use of such relations is expected to improve the performance on the retrieval and classification task. Furthermore, when one data type is missing, the correlation between data of different types allows for inference of features of the missing types from the observed types. In this work, we use the probabilistic graphical model framework to explicitly model the joint statistical association between different modalities. In particular, our effort is focused on modeling TV show data where closed-captions (text), audio, and video are available.

We present Multi-modal Latent Dirichlet Allocation (MM-LDA) model to learn the joint statistical structures of video, audio, and text. We introduce a unified representation for the 3 modalities by adopting the bag-of-words representation for video and audio documents. Unlike text, words in video and audio need to be identified and extracted using clustering of features. MM-LDA builds on the basic LDA model [2], which uses hidden variables, loosely termed topics, to cluster words and model word co-occurrences in text documents. Here, the hidden topics learned by MM-LDA describe how text words co-occur with video and audio words in multimedia clips. Our model can be compared to the work in [3], which extends LDA to learn the association between image regions and annotation words. However, [3] lacks the concept of words in their image region representation, preventing the model from capturing correlation between different image regions. We demonstrate the power of the representation learned by MM-LDA in 3 challenging tasks: TV clip classification, content-based retrieval, and automatic annotation.

## 2. MULTI-MODAL LATENT DIRICHLET ALLOCATION (MM-LDA)

### 2.1. Data Representation

We propose to unify the representation for the 3 data modalities by adopting the bag-of-word representation for video and audio documents. In such a representation, a document is treated as a collection of words (word ordering is ignored)

and is represented simply as a histogram of word frequency. A multimedia document consisting of text, audio, and video is summarized in 3 vectors of word counts. For text, we simply count how many of word $t$ from a dictionary of $T_t$ words occurs in each document. For video and audio data, such a concept of *word* need to first be identified and extracted via clustering of visual and acoustic features. A video clip is modeled as a collection of local spatio-temporal blocks of size 8×8×8, where each block is assigned to a codeword from a large dictionary of $T_v$ video words learned from the data. In a similar fashion, an 8-kHz audio clip is represented as a collection of 256-sample frames, where each frame is again assigned to a codeword from a dictionary of $T_a$ audio words. By tallying the number of word occurrences in a multimedia clip, we obtain the corresponding 3 vectors of word counts.

We adopt the following notation and terminology. A word is denoted using a unit-basis vector $w$ of size $T$ with exactly one non-zero entry representing the membership to only one word in a dictionary of $T$ words. We use $w[\tau]$ to denote the $\tau$ component of vector $w$. A document is a collection of $N$ word occurrences denoted by $\mathbf{w} = \{w_1, ..., w_N\}$. Our multi-modal document is denoted as $\mathbf{W} = \{\mathbf{w^t}, \mathbf{w^a}, \mathbf{w^v}\}$, where $\mathbf{w}^t = \{w_1^t, ..., w_{N_t}^t\}$ corresponds to $N_t$ occurrences of text words; $\mathbf{w}^a = \{w_1^a, ..., w_{N_a}^a\}$ corresponds to $N_a$ audio frames; $\mathbf{w}^v = \{w_1^v, ..., w_{N_v}^v\}$ corresponds to $N_v$ video blocks. A corpus of $M$ multimedia documents is denoted as $D = \{\mathbf{W}_1, .., \mathbf{W}_M\}$.

### 2.2. Model

The basic idea in LDA is the assumption that a document, as represented by a vector of word counts, is modeled as a weighted mixture of K hidden topics, where each topic can be thought as representing a basis pattern of word co-occurrences (word distribution). Multi-modal LDA builds on the basic LDA model by postulating that the same hidden topic also captures co-occurrences of words from different types (each topic now describes a pattern of how certain text words co-occur with certain audio and video features). To generate a multimedia document under MM-LDA, one needs only specify the different proportion of the K topics, denoted by $\theta$, the document contains. The topics, in turn, govern the probability of generating each audio, video, and text word. More formally, $\theta$ is sampled from a Dirichlet distribution with parameter $\alpha$. To generate each text word, we choose the hidden topic $k$ with probability Mult($\theta_k$) and choose the word $t'$ under topic $k$ with probability $\beta_{kt'}^t$. For audio and video word, we follow a similar process but instead we choose the word $t'$ with probability $\beta_{kt'}^a$ and $\beta_{kt'}^v$. A corpus of $M$ documents is generated by repeating the above process for each document. The graphical model of MM-LDA is shown in Figure 1.

#### *2.2.1. Inference*

Given the parameters of the model ($\alpha$, $\beta^t$, $\beta^a$, $\beta^v$), for each document $\mathbf{W} = \{\mathbf{w}^t, \mathbf{w}^a, \mathbf{w}^v\}$, we infer the posterior proba-

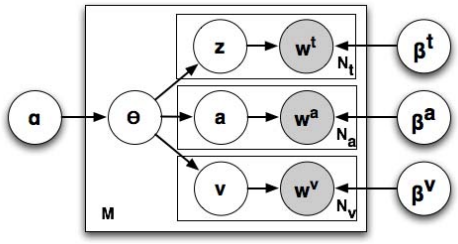


**Fig. 1**. Graphical model of the proposed Multi-modal Latent Dirichlet Allocation (MM-LDA) model.

bility over the hidden topics and $\theta$ given the observed words $p(\mathbf{z}, \mathbf{a}, \mathbf{v}, \theta|\mathbf{W})$, where $\mathbf{z} = \{z_1, ..., z_{N_t}\}$, $\mathbf{a} = \{a_1, ..., a_{N_a}\}$, and $\mathbf{v} = \{v_1, ..., v_{N_v}\}$. Similar to the case in LDA, computing the joint posterior distribution over the hidden nodes and parameter in our model is intractable. We employ the variational Bayesian framework in [4] and approximate the joint posterior with a variational posterior in a factorized form: $p(\mathbf{z}, \mathbf{a}, \mathbf{v}, \theta|\mathbf{W}) \approx q(\mathbf{z}, \mathbf{a}, \mathbf{v}|\mathbf{W})\, q(\theta|\mathbf{W})$. The problem now becomes one of finding such a variational posterior that minimizes the KL distance between the true and the factorized posterior. By differentiating the KL distance w.r.t. $q(\mathbf{a}, \mathbf{z}, \mathbf{v})$ and set the derivative to 0, we obtain the following:

$$\log q(\mathbf{z}, \mathbf{a}, \mathbf{v}) = \sum_i^{N_t} \log q(z_i) + \sum_j^{N_a} \log q(a_j) + \sum_n^{N_v} \log q(v_n)$$

$$q(z_i) \propto \exp\left( \int \left[\log p(w_i^t|z_i) + \log p(z_i|\theta)\right] q(\theta) d\theta \right)$$

$$q(z_i = k) \propto \exp\left( \sum_{t'}^{T_t} w_i^t[t'] \log \beta_{kt'}^t + E_{q(\theta)}[\log \theta_k] \right) \quad (1)$$

where $q(z_i)$, $q(a_j)$, $q(v_n)$ are short for $q(z_i|w_i^t)$, $q(a_j|w_j^a)$, and $q(v_n|w_n^v)$. The variational posterior over the hidden topics for audio and video words $q(a_j = k)$ and $q(v_n = k)$ can be computed in the same manner as in Eqn 1. The variational posterior distribution over the parameter $\theta$ can be computed from:

$$\log q(\theta) \propto \sum_{\mathbf{z},\mathbf{a},\mathbf{v}} \log p(\mathbf{W}, \mathbf{z}, \mathbf{a}, \mathbf{v}|\theta) q(\mathbf{z}) q(\mathbf{a}) q(\mathbf{v}) + \log p(\theta|\alpha)$$

$$= \sum_k \left[ \sum_i^{N_t} q(z_i) + \sum_j^{N_a} q(a_j) + \sum_n^{N_v} q(v_n) + \alpha_k - 1 \right] \log \theta_k + \ldots$$

Since the prior $p(\theta)$ is chosen to be Dirichlet, which is the conjugate of the likelihood term above, we find that the variational posterior $q(\theta)$ must also be Dirichlet. By denoting $q(\theta) \sim \text{Dir}(\tilde{\alpha})$, we obtain the following update rule:

$$\tilde{\alpha}_k = \sum_i^{N_t} q(z_i) + \sum_j^{N_a} q(a_j) + \sum_n^{N_v} q(v_n) + \alpha_k. \quad (2)$$

Since the posterior over $\theta$ takes for the form of a Dirichlet distribution, now we can compute the expectation term $E[\log \theta_k]$ in Eqn 1 to be $E[\log \theta_k] = \Psi(\tilde{\alpha}_k) - \Psi(\sum_{k'} \tilde{\alpha}_{k'})$, where $\Psi$ is the digamma function.

#### 2.2.2. Parameter Estimation

Given a corpus of documents $D = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$, we use the variational EM framework to learn the model parameters $\{\alpha, \beta^t, \beta^a, \beta^v\}$. Variational EM alternates between inferring the variational posterior probability in the E-step and finding parameter updates that maximize the lower bound of the log likelihood (M-step). Replacing the true posterior with variational posterior, we obtain the lower bound of the log likelihood $F = \sum_m F_m$ given below.

$$F_m = E_q\left[\log P(\mathbf{W}_m, \mathbf{z}_m, \mathbf{a}_m, \mathbf{v}_m, \theta_m)\right] + H(q) \qquad (3)$$

where the expectation in Eqn 3 is taken w.r.t the variational posterior $q(\mathbf{z}_m, \mathbf{a}_m, \mathbf{v}_m, \theta_m) = q(\mathbf{z}_m)q(\mathbf{a}_m)q(\mathbf{v}_m)q(\theta_m)$. The closed-form update rule for $\beta^t$ is obtained below by setting the derivative $\frac{\partial F}{\partial \beta^t_{kt'}}$ to 0. Similar update rules can be derived for $\beta^a$ and $\beta^v$.

$$\beta^t_{kt'} = \frac{\sum_{m,i} q(z_{mi} = k) w^t_{mi}[t']}{\sum_{k,m,i} q(z_{mi} = k) w^t_{mi}[t']} \qquad (4)$$

To update the Dirichlet parameter $\alpha$, since no closed-form solution is available, we use the following gradient ascent rule:

$$\Delta\alpha_k = \Psi(\sum_{k'} \alpha_{k'}) - \Psi(\alpha_k) + \frac{1}{M}\sum_m E_{q(\theta)}[\log\theta_{mk}]$$
$$\alpha^{new} = \alpha^{old} + \gamma\Delta\alpha_k$$

where $E_{q(\theta)}[\log\theta_{mk}] = \Psi(\tilde{\alpha}_{mk}) - \Psi(\sum_k \tilde{\alpha}_{mk})$ and $\gamma$ denotes the learning rate.

#### 2.2.3. Audio and Video words

While there exists a great variety of audio and video features to choose from in the related literature of speech/audio and image/video processing, we focus our attention on a few features that have been known to be discriminant in the tasks of signal classification and recognition.

*Audio words:* Short-term spectral features have been used extensively to represent audio signals, especially human speech. In this work, 8-kHz input audio signals are first divided into frames of size 256 samples. 256-pt FFT is then performed on each frame and the log of magnitude spectrum are used as our spectral features. A dictionary of audio words is then learned by fitting a mixture of Gaussians model with diagonal covariance to a collection of spectral features from the training data. The number of Gaussians used corresponds to the desired number of audio words in the dictionary.

*Video words:* In video processing, motion has long been used in a variety of tasks ranging from object segmentation, activity recognition, to video compression. In this work, we adopt the motion representation from our previous work in [5]. Each input video is first divided into blocks of size $8\times8\times8$. Spatio-temporal ICA filters resembling moving gabor filters are then used to map input pixels to ICA coefficients which are used as our motion features. A dictionary of video words are learned by fitting a mixture of Laplacians model to a collection of ICA coefficients from the training data.

## 3. EXPERIMENTS

### 3.1. Recorded TV Show Data

A database of TV shows have been recorded using a PC that runs Windows Media Center™ personal video recorder. Most recorded shows are either 30-min or 1-hr long. Because of a large amount of video and audio data corresponding to each TV show, a document is defined as a 20-sec segment (clip). For a 1-hr TV show, for example, we obtain approximately 180 clips. In our current implementation, no pre-processing has been done to remove the commercials. Therefore, approximately 10-15% of 180 clips generated from a 1-hr TV show would correspond solely to commercials.

### 3.2. TV Clip Classification

To demonstrate the power of our model in representing multi-type documents, we use MM-LDA in a 6-class classification task. We define 6 classes from the following 6 TV shows: "$40-a-Day", "CSI", "Good Eats", "Law&Order", "Modern Marvels", and "The West Wing". In order to have roughly the same number of documents (clips) in each class, we use 6 episodes of 30-min shows ("$40-a-Day" and Good Eats) and 3 episodes of 1-hr shows ("CSI", "Law&Order", "Modern Marvels", and "The West Wing") to create a 2267-document data set used in our experiments. We learn the MM-LDA model with 10 hidden topics for each class. To classify a new clip, we compute the lower bound of the log likelihood (as shown in Eqn 3) of the new clip under each of the 6 models. The clip is then classified to the class that yields the highest log likelihood. We compare the classification results from using a single modality alone to the results using 2 and 3 modalities together. 4-fold cross validation is used and the results are averaged. Figure 2 shows the performance gain when using audio and video words together. In the class where audio performs poorly, e.g. class 4–Law&Order, the video portion of the document is able to compensate and allows for improved classification results. Using audio and video words together can be interpreted as expanding the original set of vocabulary to include all possible combinations of audio-video word pairs. The use of a much larger set of vocabulary thus explains the boosted performance.

Table 1 summarizes the classification results when varying the numbers of audio and video words. In general, more words allow for better classification results. We note that using texts from closed captions alone already yields good performance, but adding audio and video words allow the results to further improve.

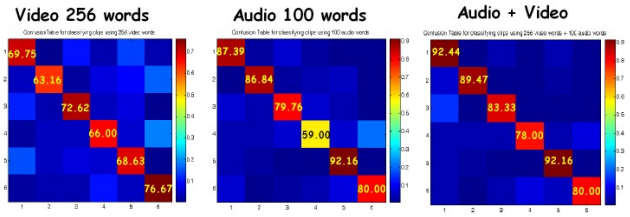


**Fig. 2**. Confusion Table for 6-class classification using 256 video words(left), 100 audio words(middle), and the combined audio and video words(right). Video words complement audio words to improve the classification performance .

**Table 1**. Summary of classification results.

| | Accuracy |
|---|---|
| 64 Audio words | 77.28% |
| 100 Audio words | 80.75% |
| 256 Video words | 68.3% |
| 512 Video words | 69.49% |
| 11206 Text words | 93.5% |
| 100 Audio words + 256 Video words | 87.78% |
| 100 Audio words + 256 Video words + Text | 94.26% |

### 3.3. Multimedia Retrieval

For the retrieval task, a mixture of MM-LDA model is used, where MM-LDA is learned for each class of TV shows in the database. A query TV clip is first classified as belonging to one of the classes (denoted by $\hat{c}$) and the variational posterior over the hidden variable $\theta$ under class $\hat{c}$ : $q(\theta|\mathbf{W}_{\text{query}}, \alpha_{\hat{c}}, \beta_{\hat{c}})$ is computed. Given our MM-LDA hidden variable model, a natural choice of distance metric between the query and a clip in the dataset is the KL distance between the posterior over the hidden variables $\theta$ for the query, and that posterior for the clip. Since the variational posterior is Dirichlet, the KL distance can be calculated analytically.

### 3.4. Automatic Annotation

Automatic annotation is a task of inferring texts from a TV clip that contains no closed captions. The idea is to infer the variational posterior of $\theta$ from the video-audio portion of the document and use the inferred posterior to predict the most likely closed caption words using the model parameters learned from a corpus of TV clips containing closed captions.

$$p(w^t|\mathbf{w^a}, \mathbf{w^v}) \approx \int \sum_{z=k} p(w^t|z)p(z|\theta)q(\theta|\mathbf{w^a}, \mathbf{w^v})d\theta$$
$$= \sum_k \beta_k^t E[\theta_k]$$

where the variational posterior $q(\theta|\mathbf{w^a}, \mathbf{w^v}) \sim \text{Dir}(\tilde{\alpha})$ and $E[\theta_k] = \frac{\tilde{\alpha}_k}{\sum_{k'} \tilde{\alpha}_{k'}}$. Figure 3 shows an annotation example of a clip from the show $40-a-day (from the Food Networks channel). The word *Rachael* is predicted, for example, because the host of the show is named Rachael Ray.

**Fig. 3**. Original Closed-captions: "the paper also mentioned that a local favorite breakfast is key lime pie. It seems a little odd to me that people eat pie for breakfast, but with the abundance of lime trees here in the Keys, I'm willing to bet that the price for a slice of pie..." Predicted words from video and audio: *'Rachael', 'like', 'gonna', 'yeah', key', 'right', 'eat', 'know', 'fun', 'just', 'good', 'great', 'food', 'fresh', 'time', 'town', 'local', 'lot', 'breakfast', 'thank', 'lime', 'make', 'best', 'cream', 'Keys', 'looks', 'affordable', 'come', 'prices'*

## 4. FUTURE EXTENSIONS

For future work, we would like to explore more choices of features, i.e. MFCCs for audio words, and color for video words. Whereas the model presented here assumes i.i.d data, actual multimedia clips are characterized by strong temporal correlations. Modeling such correlations is expected to enhance performance on the tasks we have discussed. By analogy, it is well known that incorporating language models significantly enhances the performance of speech recognition systems. We are currently extending our model to incorporate relevant multi-modal temporal statistics.

## 5. REFERENCES

[1] T. Westerveld, T. Ianeva, L. Boldareva, A. de Vries, and D. Hiemstra, "Combining information sources for video retrieval," TRECVID 2003 Workshop, 2004.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] D. M. Blei and M. I. Jordan, "Modeling annotated data," ACM Int. Conf. on Research and Development in Information Retrieval, 2003.

[4] H. Attias, "A variational bayesian framework for graphical models," Advances in Neural Information Processing Systems, 2000, vol. 12.

[5] D. Putthividhya and T.-W. Lee, "Motion patterns: High-level representation of natural video sequences," IEEE Conf. on Computer Vision and Pattern Recognition, 2006, vol. 12.