

Harmonium Models for Semantic Video Representation and Classification

Jun Yang Yan Liu Eric P. Xing Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
{juny, yanliu, epxing, alex}@cs.cmu.edu

Abstract

Accurate and efficient video classification demands the fusion of multimodal information and the use of intermediate representations. Combining the two ideas into the same framework, we propose a probabilistic approach for video classification using intermediate semantic representations derived from the multi-modal features. Based on a class of bipartite undirected graphical models named *harmonium*, our approach represents video data as latent semantic topics derived by jointly modeling the transcript keywords and color-histogram features, and perform classification using these latent topics under a unified framework. We show satisfactory classification performance of our approach on a benchmark dataset, and some interesting insights of the data provided by this approach.

1 Introduction

Classifying video data into semantic categories, sometimes known as semantic video concept detection, is an important research topic. Video data contain multifarious data types including image frames, transcript text, speech, audio signal, each bearing correlated and complementary information essential to the analysis and retrieval of video data. The fusion of such multimodal information is regarded as a key research problem [10], and has been a widely used technique in video classification and retrieval methods. Many fusion strategies have been proposed, varying from early fusion [12], which merges the feature vectors extracted from different modalities, to late fusion, which combines the outputs of the classifiers or “retrieval experts” built on each single modality [12, 6, 18, 15]. Empirical results show that methods based on the fusion of multimodal information outperforms those based on any single type of information in both video classification and retrieval tasks.

Another trend in video classification is the seek of low-dimensional, intermediate representations of video

data. The primary reason is to make sophisticated classifiers (e.g, SVM) affordable, which otherwise would be computationally expensive on the high-dimensional raw features. Moreover, using intermediate representations holds the promise of better interpretation of the data semantics, and may lead to superior classification performance. Work along this direction includes the conventional dimension-reduction methods such as principal component analysis (PCA) and Fisher linear discriminant (FLD) [4], as well as probabilistic methods such as probabilistic latent semantic indexing (pLSI) [5], latent Dirichlet allocation (LDA) [2], exponential-family harmonium (EFH) [14]. While many of the models are for single-modal data such as textual documents, there are also extensions for modeling multi-modal data such as captioned images and video [1, 17].

The key insights for video classification from previous works appear to be combining multimodal information and using intermediate representations. The goal of this paper is to take advantage of both insights using an integrated and principled approach. Based on a class of bipartite undirected graphical models (i.e., random fields) called *harmonium* [14, 17], our approach extracts intermediate representation as *latent semantic topics* of video data by jointly modeling the correlated information in image regions and transcript keywords. Moreover, this approach explicitly introduces category label(s) into the model, which allows the classification and representation to be accomplished in a unified framework.

The proposed approach is significantly different from previous models for text/multimedia data, mainly in that it incorporates category labels as (hidden) model variables, in addition to the variables representing data (features) and latent semantic topics. This allows the classification to be done by inferencing the distribution of these label variables conditioned on the observed data variables. In contrast, existing models [2, 1, 5, 14, 17] are mainly for deriving the intermediate data repre-

sentation in the form of latent semantic topics. The classification, if needed, is performed using a separate classifier built on top of the derived intermediate representations. Therefore, one advantage of our approach is the unified model for both representation and classification, which avoids the need of building separate classifiers. More importantly, by modeling the interactions between latent semantic topics and category labels, this approach tailors the intermediate representation to reflect category information of the data. Such “supervised” intermediate representations are expected to provide more discriminative power and insights of the data than the “unsupervised” representations generated by existing methods [2, 1, 5, 14, 17].

Our proposal includes two related models, each bearing different implications to the representation and classification of the video data. *Family-of-harmonium* (FoH) builds a family of category-specific harmonium models, with each one modeling the video data from a specific category. The label of a video shot is predicted by comparing its likelihood in each harmonium model. *Hierarchical harmonium* (HH) incorporates all the category labels as an additional layer of hidden variables into a single harmonium model, and performs classification through the inference of these label variables. Due to the model structure, the FoH model reveals the internal structure of each category, and can be easily extended to include new categories without retraining the whole model. In contrast, the HH model reveals the relationships between multiple categories, and takes advantage of such relationships in classification.

In Section 2 we review the related work on the fusion of multimodal video features as well as representation models for video data. We describe the two proposed models in Section 3. In Section 5, we show that the proposed models achieve satisfactory classification performance and interesting data interpretation through experiments on TRECVID video collection. The conclusions and future work are discussed in Section 6.

2 Related Works

As pointed out in [10], the processing, indexing, and fusion of data in multiple modalities is a core problem of multimedia research. For video classification and retrieval, the fusion of features from multiple data types (e.g., key-frames, audio, transcript) allows them to complement each other to achieve better performance than using any single type of feature, and has been widely used in many existing methods. The fusion strategies vary from early fusion [12], which merges the feature vectors extracted from different data modalities, to late fusion, which combines the output of classifiers or “retrieval experts” built on each single modality [12, 6,

18, 15]. It remains an open question as to which fusion strategy is more appropriate for a certain task, and a comparison of the two strategies in video classification is presented in [12]. The approach to be presented here uses neither of these two fusion strategies; instead, it derives the latent semantic representation of the video data by jointly modeling the multimodal low-level features, so the fusion takes place somewhere between early fusion and late fusion.

There are many approaches to obtain low-dimensional intermediate representations of video data. Principal component analysis (PCA) has been the most popular method, which projects the raw features into a lower-dimensional feature space where the data variances are well preserved. Independent component analysis (ICA) and Fisher linear discriminant (FLD) are the other popular dimension reduction methods. Recently, there are also many proposals on modeling the latent semantic topics of the text and multimedia data. For example, latent semantic indexing (LSI) by Deerwester et al. [3] transforms term counts linearly into a low-dimensional semantic eigenspace, and the idea was later extended by Hofmann to probabilistic LSI (pLSI) [5]. The latent Dirichlet allocation (LDA) by Blei et al. [2] is a directed graphical model that provides generative semantics of text documents, where each document is associated with a topic-mixing vector and each word is independently sampled according to a topic drawn from this topic-mixing. LDA has been extended to Gaussian-Mixture LDA (GM-LDA) and Correspondence LDA (Corr-LDA) [1], both of which are used to model annotated data such as captioned images or video with transcript text. Exponential-family harmonium (EFH) proposed by Welling et al. [14] is bipartite undirected graphical model consisting a layer of latent nodes representing semantic aspects and a layer of observed nodes representing raw data (features). To model multi-modal data, Eric et al. [17] has extended it to the multi-wing harmonium model where the data layer consists of two or more “wings” of nodes representing textual, imagery, and other types of data, respectively.

In practice, the methods mentioned above are mainly used for transforming the high-dimensional raw data into a low-dimensional representation which presumably capture the latent semantics of the data. Classification task is usually performed by building a separate discriminative classifier (e.g., SVM) based on such latent semantic representations. In this paper, we seek for a more unified approach where the representation and classification can be integrated into the same framework. This approach not only achieves satisfactory classification performance, but also provides interesting in-

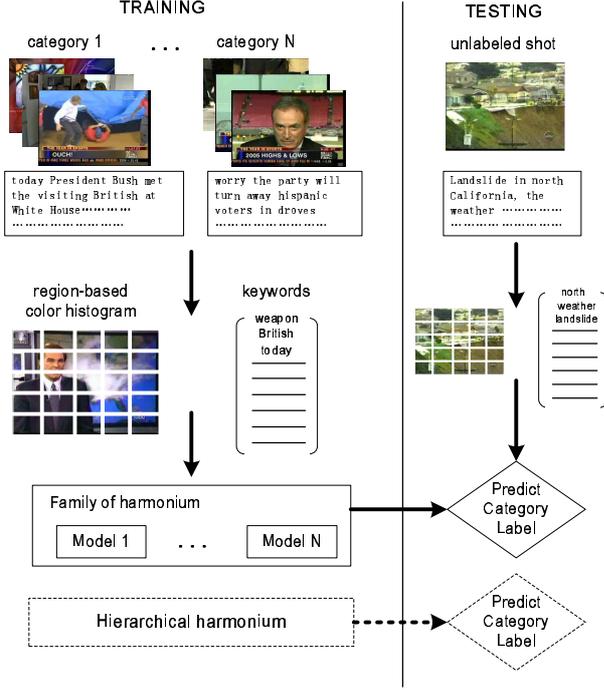


Figure 1: A sketch of our approach

sights into the data semantics, such as the internal structure of each category and the relationships between different categories. Fei-Fei et al. [8] used a unified model for representing and classifying natural scene images by introducing category variables into the LDA model, which is similar to our approach except that our models are undirected.

3 Our Approach

A sketch of our approach is illustrated in Figure 1. The basic data unit to be classified is called video shots, namely video segments with length varying from a few seconds to half minute or even longer. We represent each video shot by a bag of keywords extracted from the video transcript, which is obtained from the video’s closed-captions or speech recognition, as well as the a set of fixed-sized image regions extracted from the keyframe of the video shot. Each region is described by its color histogram feature. We learn the model that best describes the joint distribution of the keywords and color features of the video shots in each category. In classification, we extract the keywords and color features from an unlabeled video shot, from which we derive the most likely category this shot belongs to. The two proposed models, family-of-harmonium and hierarchical harmonium, differ in the way the data are modeled and classified.

Both of our models are based on a class of bipar-

tite undirected model (i.e., random fields) called *harmonium*, which has been used by Welling et al. [14] and Eric et al. [17] to model text and multimedia data. Our models use their models as the basic building block, but differ from theirs by explicitly incorporating the category labels into the model. This allows our models to represent and classify video data in a unified framework, while the previous harmonium models are only for data representation.

3.1 Notations and definitions The notations used in the paper follow the convention of probabilistic models. Uppercase characters represent random variables, while lowercase characters represent the instances (values) of the random variables. Bold font is used to indicate a vector of random variables or their values. In the illustrations, shaded circles represent observed nodes while unfilled circles represent hidden (latent) nodes. Each node in a graphical model is associated with a random variable, and we use the term node and variable interchangeably in this paper.

The semantics of the model variables are described as follows:

- A video shot s is represented by a tuple as $(\mathbf{x}, \mathbf{z}, \mathbf{h}, \mathbf{y})$, which respectively denote the keywords, region-based color features, latent semantic topics, and category labels of the shot.
- The vector $\mathbf{x} = (x_1, \dots, x_N)$ denotes the keyword feature extracted from the transcript associated with the shot. Here N is the size of the word vocabulary, and $x_i \in \{0, 1\}$ is a binary variable that indicates the absence or presence of the i^{th} keyword (of the vocabulary) in the shot.
- The vector $\mathbf{z} = (z_1, \dots, z_M)$ denotes color-histogram features of the keyframe of the shot. Each keyframe is evenly divided into a grid of totally M fixed-sized rectangular regions, and $z_j \in \mathcal{R}^C$ is a C -dimensional vector that represents the color histogram of the j^{th} region. So \mathbf{z} is a stacked vector of length equal to CM .
- The vector $\mathbf{h} = (h_1, \dots, h_K)$ represents the latent semantic topics of the shot, where K is the total number of the latent topics. Each component $h_k \in \mathcal{R}$ denotes how strongly this shot is associated with the k^{th} latent topic.
- The category labels of a shot are modeled differently in the two models. In family-of-harmonium, a single variable $y \in \{1, \dots, T\}$ indicates the category this shot belongs to, where T is the total number of categories. In hierarchical harmonium, the

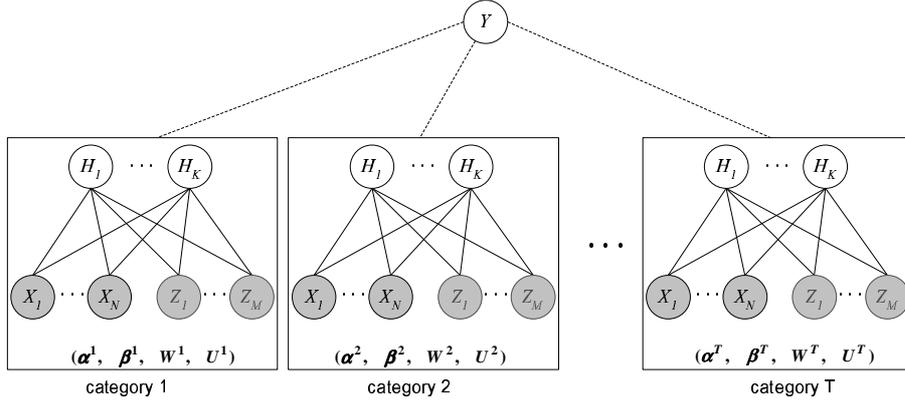


Figure 2: The family-of-harmonium model

labels are represented by a vector $\mathbf{y} = (y_1, \dots, y_T)$, with each $y_t \in \{0, 1\}$ denoting whether the shot is in the t^{th} category. Here a video shot belongs to only one category, so we have $\sum_t y_t = 1$.

- The two proposed models have different sets of parameters. The family-of-harmonium has a specific harmonium model for each category y , with parameters as $\theta^y = (\pi_y, \alpha^y, \beta^y, W^y, U^y)$. The hierarchical harmonium has a single set of parameters as $\theta = (\alpha, \beta, \tau, W, U, V)$.

3.2 Family-of-harmonium (FoH) The FoH model is illustrated in Figure 2. It contains a set of T category-specific harmoniums, with each harmonium modeling the video data from a specific category. Each harmonium is a bipartite undirected graphical model that consists of two layers of nodes. Nodes in the top layer represent the *latent* semantic topics $\mathbf{H} = \{H_k\}$ of the data. To represent the bi-modal features of video data, the bottom layer contains two “wings” of *observed* nodes that represent the keyword features $\mathbf{X} = \{X_i\}$ and region-based color features $\mathbf{Z} = \{Z_j\}$, respectively. Each node is linked with all the nodes in the opposite layer, but not with any node in the same layer. This topology ensures that the nodes in one layer are *conditionally independent* given the nodes in the opposite layer, a property important to the construction and inference of the model. All the component harmoniums in FoH share exactly the same structure, but have unique set of parameters $\theta^y = (\pi_y, \alpha^y, \beta^y, W^y, U^y)$ indexed by the category label y .

We now describe the distributions of these variables. The category label Y follows a prior distribution as a multinomial:

$$(3.1) \quad p(y) = \text{Multi}(\pi_1, \dots, \pi_T),$$

where $\sum_{t=1}^T \pi_t = 1$. In FoH, Y is not actually linked with any nodes in the component harmoniums; instead, it serves as an *indicator variable* that selects a specific harmonium from the whole family of harmoniums to model the video data of a particular category. In the distribution function of each harmonium, Y only appears as the subscript of the model parameters.

Given its category label y , we see the raw features of a shot as well as its latent semantic topics as two layers of representations mutually influencing each other in the specific harmonium associated with this category. We can either conceive keyword and color features as being generated by the latent semantic topics, or conceive the semantic topics as being summarized from the keyword and image features. This mutual influence is reflected in the conditional distributions of the variables representing the features and the semantic topics.

For the keyword feature, the variable x_i indicating the presence/absence of term $i \in \{1, \dots, N\}$ in the vocabulary is distributed as:

$$(3.2) \quad P(X_i = 1 | \mathbf{h}, y) = \frac{1}{1 + \exp(-\alpha_i^y - \sum_k W_{ik}^y h_k)}$$

$$P(X_i = 0 | \mathbf{h}, y) = 1 - P(X_i = 1 | \mathbf{h}, y)$$

This shows that the each keyword in a video shot is sampled from a Bernoulli distribution dependent on the latent semantic topics \mathbf{h} . That is, the probability that a keyword appears is affected by a weighted combination of semantic topics \mathbf{h} . Parameter α_i^y and W_{ik}^y are both scalars, so $\alpha^y = (\alpha_1^y, \dots, \alpha_N^y)$ is a N -dimensional vector, and $W^y = [W_{ik}^y]$ is a matrix of size $N \times K$. Due to the conditional independence between x_i given \mathbf{h} , we have $p(\mathbf{x} | \mathbf{h}, y) = \prod_i p(x_i | \mathbf{h}, y)$.

The color-histogram feature z_j of the j^{th} region in the keyframe of the shot admits a conditional multivari-

ate Gaussian distribution as:

$$(3.3) \quad p(z_j|\mathbf{h}, y) = \mathcal{N}(z_j|\Sigma_j^y(\beta_j^y + \sum_k U_{jk}^y h_k), \Sigma_j^y)$$

where z_j is sampled from a distribution parameterized by the latent semantic topics \mathbf{h} . Here, both β_j^y and U_{jk}^y are C -dimensional vectors, and therefore $\beta^y = (\beta_1^y, \dots, \beta_M^y)$ is a stacked vector of dimension CM and $U^y = [U_{jk}^y]$ is a matrix of size $CM \times K$. Note that Σ_j^y is a $C \times C$ covariance matrix, which, for simplicity, is set to identity matrix I in our model. Again, we have $p(\mathbf{z}|\mathbf{h}, y) = \prod_j p(z_j|\mathbf{h}, y)$ due to conditional independence.

Finally, each latent topic variable h_j follows a conditional univariate Gaussian distribution whose mean is determined by a weighted combination of the keyword feature \mathbf{x} and the color feature \mathbf{z} :

$$(3.4) \quad p(h_k|\mathbf{x}, \mathbf{z}, c) = \mathcal{N}(h_k|\sum_i W_{ik}^y x_i + \sum_j U_{jk}^y z_j, 1)$$

where W_{ik}^y and U_{jk}^y are the same parameters used in Eq.(3.2) and (3.3). Similarly, $p(\mathbf{h}|\mathbf{x}, \mathbf{z}, y) = \prod_k p(h_k|\mathbf{x}, \mathbf{z}, y)$ holds.

So far we have presented the conditional distributions of all the variables in the model. These local conditionals can be mapped to the following harmonium random fields as:

$$(3.5) p(\mathbf{x}, \mathbf{z}, \mathbf{h}|y) \propto \exp \left\{ \sum_i \alpha_i^y x_i + \sum_j \beta_j^y z_j - \sum_j \frac{z_j^2}{2} - \sum_k \frac{h_k^2}{2} + \sum_{ik} W_{ik}^y x_i h_k + \sum_{jk} U_{jk}^y z_j h_k \right\}$$

We present the details of the derivation of this random field in the Appendix. Note that the partition function (global normalization term) of this distribution is not explicitly shown, so we use proportional sign instead of equal sign. This hidden partition function increases the difficulty of learning the model.

By integrating out the hidden variables \mathbf{h} in Eq.(3.5), we obtain the category-conditional distribution over the observed keyword and color features of a video shot:

$$(3.6) p(\mathbf{x}, \mathbf{z}|y) \propto \exp \left\{ \sum_i \alpha_i^y x_i + \sum_j \beta_j^y z_j - \sum_j \frac{z_j^2}{2} + \frac{1}{2} \sum_k (\sum_i W_{ik}^y x_i + \sum_j U_{jk}^y z_j)^2 \right\}.$$

There is also a hidden partition function in this distribution. The marginal distribution (likelihood) of a labeled video shot can be decomposed into a category-specific marginal and a prior over the categories, i.e., $p(\mathbf{x}, \mathbf{z}, y) = p(\mathbf{x}, \mathbf{z}|y)p(y)$.

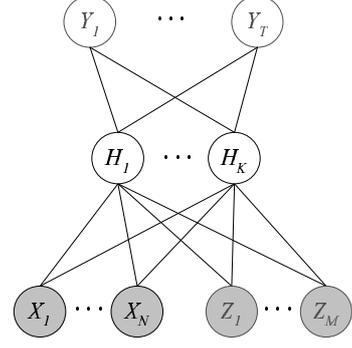


Figure 3: Hierarchical harmonium model

The learning of FoH involves learning T component harmoniums, with each harmonium learned independently using the (labeled) video shots from the corresponding category. To learn the harmonium model for a category y , we estimate its model parameters $\theta^y = (\alpha^y, \beta^y, W^y, U^y)$ by maximizing the likelihood of the video shots in category y , where the likelihood function is defined by Eq.(3.6). Due to the existence of partition function, the learning requires approximate inference methods. We will further discuss the learning methods in Section 4.

The category of an unlabeled shot is predicted by finding the component harmonium that best describes the features of the shot. Given the keyword feature \mathbf{x} and color feature \mathbf{z} of a shot, we compute the posterior probability of each category label as:

$$(3.7) \quad p(y|\mathbf{x}, \mathbf{z}) \propto p(\mathbf{x}, \mathbf{z}|y)p(y) \propto p(\mathbf{x}, \mathbf{z}|y)$$

The second step in the derivation assumes that the category prior is a uniform distribution as $p(y) = 1/T$. Eq.(3.7) indicates that we can predict the category of a shot by comparing its likelihood $p(\mathbf{x}, \mathbf{z}|y)$ in each of the category-specific harmonium models computed by Eq.3.6. The harmonium that best fits the shot determines its category.

3.3 Hierarchical harmonium (HH) The second proposed model, hierarchical harmonium, adopts a different way of incorporating category labels into the basic harmonium model. Instead of building a separate harmonium for each category, it introduces the category labels as another layer of nodes $\mathbf{Y} = \{Y_1, \dots, Y_T\}$ into a single harmonium, with $Y_t \in \{0, 1\}$ indicating a shot's membership with category t . As illustrated in Figure 3, these label variables \mathbf{Y} form a bipartite subgraph with the latent topic nodes \mathbf{H} . There is a link between any Y_t and H_j but not between two Y_t , which are conditionally independent given \mathbf{H} . Unlike FoH, there is only a single hierarchical harmonium in this model.

In the HH model, the conditional distribution of \mathbf{x} and \mathbf{z} stay the same as those in the FoH model, which are defined by Eq.(3.2) and Eq.(3.3), respectively. The only difference is that the model parameters $\theta = (\alpha, \beta, \tau, W, U, V)$ no longer depend on category labels. The introduced label variable Y_t follows a Bernoulli distribution as:

$$(3.8) \quad \begin{aligned} P(Y_t = 1|\mathbf{h}) &= \frac{1}{1 + \exp(-\tau_t - \sum_k V_{tk} h_k)} \\ P(Y_t = 0|\mathbf{h}) &= 1 - P(Y_t = 1|\mathbf{h}) \end{aligned}$$

where $V = [V_{tk}]$ is a matrix of size $T \times K$. Note that if we treat \mathbf{h} as input, V_{tk} and τ as parameters, this distribution has exactly the same form as the posterior distribution of the class label in logistic regression [4], i.e., $P(Y = 1|x) = 1/(1 + \exp(-\beta_0 - \beta^T \mathbf{x}))$. This implies that the model is actually performing logistic regression to compute each category label Y_t using the latent semantic topics \mathbf{h} as input.

The distribution of each latent topic variable h_k needs to be modified to incorporate the interactions between label variables \mathbf{y} and the topic variables \mathbf{h} :

$$(3.9) \quad p(h_k|\mathbf{x}, \mathbf{z}, \mathbf{y}) = \mathcal{N}(h_k | \sum_i W_{ik} x_i + \sum_j U_{jk} z_j + \sum_t V_{tk} y_t, 1)$$

Therefore, the distribution of the latent semantic topics are not only affected by the data features \mathbf{x} and \mathbf{z} , but also by their labels \mathbf{y} . This is a significant difference from existing harmonium models [14, 17] where the distribution of latent topics only depend on the data.

With the incorporation of label variables, the random field of hierarchical harmonium becomes:

$$(3.10) \quad \begin{aligned} p(\mathbf{x}, \mathbf{z}, \mathbf{h}, \mathbf{y}) \propto \exp & \sum_i \alpha_i x_i + \sum_j \beta_j z_j - \sum_j \frac{z_j^2}{2} + \sum_t \tau_t y_t \\ & - \sum_k \frac{h_k^2}{2} + \sum_{ik} W_{ik} x_i h_k + \sum_{jk} U_{jk} z_j h_k + \sum_{tk} V_{tk} y_t h_k \end{aligned}$$

After integrating out the hidden variable \mathbf{H} , the marginal distribution of a *labeled* video shot $(\mathbf{x}, \mathbf{z}, \mathbf{y})$ is:

$$(3.11) \quad \begin{aligned} p(\mathbf{x}, \mathbf{z}, \mathbf{y}) \propto \exp & \sum_i \alpha_i x_i + \sum_j \beta_j z_j - \sum_k \frac{z_k^2}{2} + \sum_t \tau_t y_t \\ & + \frac{1}{2} \sum_k (\sum_i W_{ik} x_i + \sum_j U_{jk} z_j + \sum_t y_t V_{tk})^2 \end{aligned}$$

The parameters of the HH model, $\theta = (\alpha, \beta, \tau, W, U, V)$, are estimated under the maximum likelihood principle using the likelihood function defined

by Eq.(3.11). The classification is performed in a very different way in HH. To predict the category of an unlabeled video shot, we need to infer the unknown label variables \mathbf{Y} of the shot, from its keyword and color features. This is done by computing the conditional probability $p(Y_t = 1|\mathbf{x}, \mathbf{z})$ for each label variable Y_t . The category that gives the highest conditional probability is predicted as the category of the shot:

$$(3.12) \quad t^* = \operatorname{argmax}_t p(Y_t = 1|\mathbf{x}, \mathbf{z})$$

There is, however, no analytical solution to this conditional probability. Various approximate inference methods are available to solve this probability, as further discussed in Section 4.

3.4 Model comparison We compare our models with the existing models for text and multimedia data, including pLSI [5], LDA [2] and its variants GM-LDA and Corr-LDA [1], exponential-family harmonium [14, 17]. First of all, our models not only derive the latent semantic representation of the data but also perform classification within the same framework, while all the above models are only for data representation and require separate classifiers for the classification task. This is not necessarily an advantage of our models. But in terms of classification, ours is a more integrated approach, which presumably leads to superior performance and better data interpretation. A model similar (in spirit) to our FoH model is the Bayesian hierarchical model for scene classification proposed by Fei-fei et al. [8], except that it is based on the directed LDA model. Second, in our models the category labels “supervise” the derivation of latent semantic representation. As a result, the derived representation reflects not only the characteristics of the underlying data but also the category information. This is to be contrasted to the “unsupervised” derivation of latent semantic representation in all the other models. The third issue is the choice between directed or undirected models. The harmonium models [14, 17], including ours, are all undirected models, while the rest are directed ones. Using undirected models makes the inference easier due to conditional independence of hidden variables, but the learning is usually harder due to the global normalization term.

There are also interest contrasts between the proposed models. They first differ in terms of the semantics of the latent semantic topics derived. In FoH, each harmonium model is for a specific category, and the latent topics learned in each harmonium capture the internal structure of the data in that category, i.e., they represent the themes or data sub-clusters in that particular category. There are no correspondences between the semantic topics across different harmoniums: the first

topic in one harmonium is unrelated to the first topic in another. In contrast, HH has a single set of latent semantic topics derived from the data in various categories. These semantic topics are however different from those learned by other representation models, as they are “supervised” by the category labels and presumably contain more discriminative information. Sharing a single semantic representation also help reveals the connections and differences between multiple categories. The two models also differ in terms of scalability. FoH can easily accommodate a new category by adding another harmonium trained from the data of this new category. The existing harmoniums do not need to be changed. However, introducing a new category into HH means adding another (label) node into the model, which requires re-training of the whole model since its structure is changed.

4 Learning and inference

The parameters of our models, namely $(\alpha^y, \beta^y, W^y, U^y)$ in the FoH model and $(\alpha, \beta, \tau, W, U, V)$ in the HH model, can be estimated by maximizing the data likelihood. However, there is no closed-form solution to the parameters in complex models like ours, and therefore iterative searching algorithm has to be used. As an example, we discuss the learning and inference algorithms for the HH model. The learning and inference of each component harmonium in the FoH model can be easily derived from that.

As described in the previous section, the log-likelihood of the data under the HH model is defined by Eq.(3.11). By taking derivatives of the log-likelihood function w.r.t the parameters, we have the following gradient learning rules:

$$\begin{aligned}
\delta\alpha_i &= \langle x_i \rangle_{\tilde{p}} - \langle x_i \rangle_p \\
\delta\beta_j &= \langle z_j \rangle_{\tilde{p}} - \langle z_j \rangle_p \\
\delta\tau_t &= \langle y_t \rangle_{\tilde{p}} - \langle y_t \rangle_p \\
\delta W_{ik} &= \langle x_i h'_k \rangle_{\tilde{p}} - \langle x_i h'_k \rangle_p \\
\delta U_{jk} &= \langle z_j h'_k \rangle_{\tilde{p}} - \langle z_j h'_k \rangle_p \\
\delta V_{tk} &= \langle y_t h'_k \rangle_{\tilde{p}} - \langle y_t h'_k \rangle_p
\end{aligned}
\tag{4.13}$$

where $h'_k = \sum_i W_{ik} x_i + \sum_j U_{jk} z_j + \sum_t V_{tk} y_t$, and $\langle \cdot \rangle_{\tilde{p}}$ and $\langle \cdot \rangle_p$ denotes expectation under empirical distribution (i.e., data average) or model distribution of the harmonium, respectively. Like most undirected graphical models, there is a global normalizer term in the likelihood function of harmonium, which makes directly computing $\langle \cdot \rangle_p$ intractable. Therefore, we need approximate inference methods to approximate these model expectations $\langle \cdot \rangle_p$. We explored four methods which are

briefly discussed below. The conditional distribution of the label nodes $p(Y_t = 1 | \mathbf{x}, \mathbf{z})$ is also computed using these approximate inference methods.

4.1 Mean field approximation Mean field (MF) is a variational method that approximates the model distribution p through a factorized form as a product of marginals over clusters of variables [16]. We use the naive version of mean field, where the joint probability p is approximated by an surrogate distribution q as a product of *singleton* marginals over the variables:

$$q(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{h}) = \prod_i q(x_i | \nu_i) \prod_j q(z_j | \mu_j, I) \prod_t q(y_t | \lambda_t) \prod_k q(h_k | \gamma_k)$$

where the singleton marginals are defined as $q(x_i) \sim \text{Bernoulli}(\nu_i)$, $q(z_j) \sim N(\mu_j, I)$, $q(y_t) \sim \text{Bernoulli}(\lambda_t)$, and $q(h_k) \sim N(\gamma_k, 1)$, and $\{\nu_i, \mu_j, \lambda_t, \gamma_k\}$ are variational parameters. The variation parameters can be computed by minimizing the KL-divergence between p and q , which results in the following fixed-point updating equations i.e.

$$\begin{aligned}
\nu_i &= \sigma(\alpha_i + \sum_k W_{ik} \gamma_k) \\
\mu_j &= \beta_j + \sum_k U_{jk} \gamma_k \\
\lambda_t &= \sigma(\tau_t + \sum_k V_{tk} \gamma_k) \\
\gamma_k &= \sum_i W_{ik} \nu_i + \sum_j U_{jk} \mu_j + \sum_t V_{tk} \lambda_t
\end{aligned}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function. After the fixed-point equations converge, the surrogate distribution q is fully specified by the converged variational parameters. We replace the intractable $\langle \cdot \rangle_p$ with $\langle \cdot \rangle_q$ in Eq.(4.13), which is easy to compute from the fully factorized q . Note that after each iterative searching step in Eq.(4.13), we need to recompute the variational parameters in q since the model parameters of p have been updated.

4.2 Gibbs sampling Gibbs sampling, as a special form of the Markov chain Monte Carlo (MCMC) method, has been used widely for approximate inference in complex graphical models [7]. This method repeatedly samples variables in a particular order, with one variable at a time and conditioned on the current values of the other variables. For example in our hierarchical harmonium model, we define the sampling order as $y_1, \dots, y_T, h_1, \dots, h_K$, and then sample each y_t from the conditional distribution defined in Eq.(3.8) using the

current values of h_j , finally sample each h_j according to Eq.(3.9). After a large number of iterations (“burn-in” period), this procedure guarantees to reach an equilibrium distribution that in theory is equal to the model distribution p . Therefore, we use the empirical expectation computed using the Gibbs samples collected after the burn-in period to approximate the true expectation $\langle \cdot \rangle_p$.

4.3 Contrastive divergence Instead of doing an exact gradient ascent search using the learning rules in Eq.(4.13), we can use the contrastive divergence (CD) [13] proposed by Hinton and Welling to approximate the gradient learning rules. In each step of the gradient update, instead of computing the model expectation $\langle \cdot \rangle_p$, CD runs the Gibbs sampling for up to only a few iterations and uses the resulting distribution q to approximate the model distribution p . It has been proved that the final values of the parameters by this kind of updating will converge to the maximum likelihood estimation [13]. In our implementation, we compute $\langle \cdot \rangle_q$ from a large number of samples obtained by running only *one* step of Gibbs sampling with different initializations. Straightforwardly, CD is significantly more efficient than the Gibbs sampling method since the “burn-in” process is skipped.

4.4 The uncorrected Langevin method The uncorrected Langevin method [9] is originated from the Langevin Monte Carlo method by accepting all the proposal moves. It makes use of gradient information and resembles noisy steepest ascent to avoid local optimal. Similar to the gradient ascent, the uncorrected Langevin algorithm has the following update rule:

$$(4.14) \quad \lambda_{ij}^{\text{new}} = \lambda_{ij} + \frac{\epsilon^2}{2} \frac{\partial}{\partial \lambda_{ij}} \log p(X, \lambda) + \epsilon n_{ij}$$

where $n_{ij} \sim N(0, 1)$ and ϵ is the parameter to control the step size. Like the contrastive divergence algorithm, we use only a few iterations of Gibbs sampling to approximate the model distribution p .

5 Experiments

We evaluate the proposed models using video data from the TRECVID 2003 development set [11]. Based on the manual annotations on this set, we choose 2468 shots that belong to 15 semantic categories, which are *airplane*, *animal*, *baseball*, *basketball*, *beach*, *desert*, *fire*, *football*, *hockey*, *mountain*, *office*, *road traffic*, *skating*, *studio*, and *weather news*. Each shot belongs to only one category. The size of a category varies from 46 to 373 shots. The keywords of each shot are extracted from the video closed-captions associated with

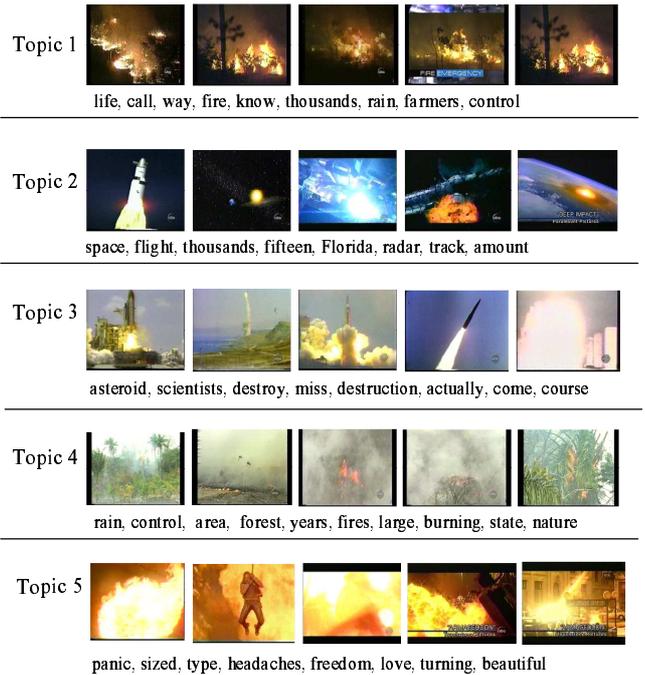


Figure 4: The representative images and keywords of 5 latent topics derived from the data in category “Fire”

that shot. By removing non-informative words such as stop words and rare words, we reduce the total number of distinct keywords (vocabulary size) to 3000. Meanwhile, we evenly divide the key-frame of each shot into a grid of 5x5 regions, and extract a 15-dimensional color histogram on HVC color space from each region. Therefore, each video shot can be represented by a 3000-d keyword feature and a 375-d color histogram feature. For simplicity, the keyword features are made binary, meaning that they only capture the presence/absence information of each keyword, because it is rare to see a keyword appears multiple times in the short duration of a shot.

The experiment results are presented in two parts. In the first part, we will show some illustrative examples of the latent semantic topics derived by the proposed models and discuss the insights they provide into the structure and relationships of video categories. In the next part, we will evaluate the performance of our models in video classification in comparison with some of the existing approaches.

5.1 Interpretation of latent semantic topics

Both the family-of-harmonium (FoH) and the hierarchical harmonium (HH) model derive latent semantic topics as intermediate representation of video data. Since each harmonium in FoH is learned independently from

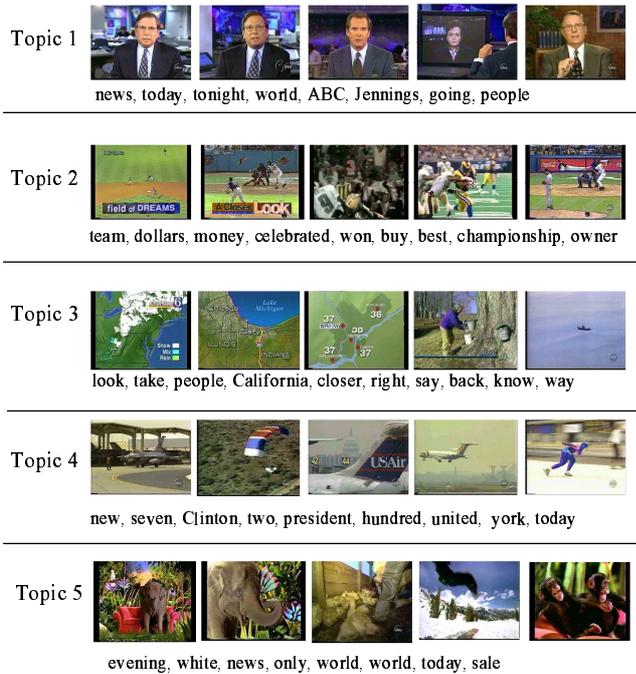


Figure 5: The representative images and keywords of 5 latent topics derived from the whole data set

the data of a specific category, its latent topics capture the structure of that particular category. To show these topics are meaningful, in Figure 4 we illustrate the 5 latent topics learned from the video category “Fire” by showing the keywords and images associated with 5 video shots that have the highest conditional probability given each latent topic. As we can see, the 5 topics roughly correspond to 5 sub-categories under the category “fire”, which can be described as “forest fire in the night”, “explosion in outer space”, “launch of missile or space shuttle”, “smoke of fire”, and “close-up scene of fire”. Since these latent topics are derived by jointly modeling the textual and image features of the video data, they are more than simply clusters in color or keyword feature space, but sort of “co-clusters” in both feature spaces. For example, the shots of Topic 1 are very similar to each other visually; the shots of Topic 2 are not so similar visually, but it is clear that they have very close semantic meanings and share common keywords such as “flight” and “radar”. The keywords associated with Topic 5 seem to be irrelevant at the first glance, but later we find that these shots contain the scenes from a movie, which explains the occurrence of keywords like “love”, “freedom”, and “beautiful”.

We also illustrate the 5 latent topics out of a set of 20 topics learned in the HH model in Figure 5. Note that these topics are learned from the whole data set instead of the data from one category, so they are expected

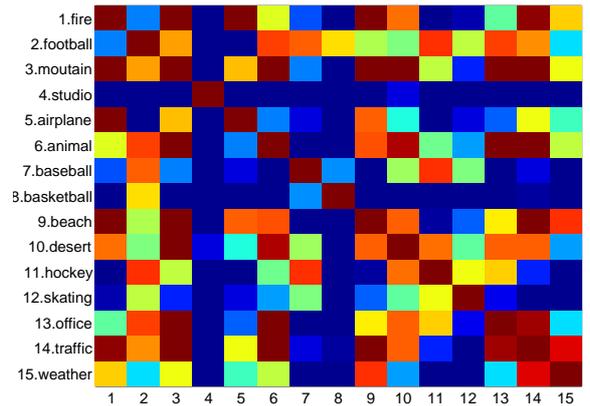


Figure 6: The color-coded matrix showing the pairwise similarity between categories. Best viewed with color.

to represent some high-level semantic topics. We can see that these 5 topics are about “studio”, “baseball or football”, “weather news”, “airplane or skating”, “animal”, which can be roughly mapped to some of the 15 categories in the data set. These results clearly show that the latent semantic topics learned by our models capture the semantics of the video data.

Another advantage of hierarchical harmonium, as we discussed in Section 3.4, is that it reveals of the relationships between different categories through the hidden topics. We can tell the how much a category t is associated with a latent topic j from the conditional probability $p(y_t|h_j)$. Therefore, we are able to compute the similarity between any two categories by examining the hidden topics they are associated with. We show the pairwise similarity between the 15 categories using the color-coded confusion matrix in Figure 6, where red(er) color denotes higher similarity and blue(er) color denotes lower similarity. We can see many meaningful pairs of related categories, e.g., “mountain” is strongly related to “animal”, “baseball” is related to “hockey”, while “studio” is not related to any category. These relationships are basically consistent with common sense.

5.2 Performance on video classification To evaluate the performance of the FoH and HH model in video classification, we evenly divide our data set into a training set and a test set. The model parameters are estimated from the training set. Specifically, we implemented the learning methods based on the four inference algorithms described in Section 4, in order to examine their efficiency and accuracy. We also explore the issue of model selection, namely the impact of the number of latent semantic topics to the classification performance.

Several other methods have been implemented for comparison, all of which produce intermediate repre-

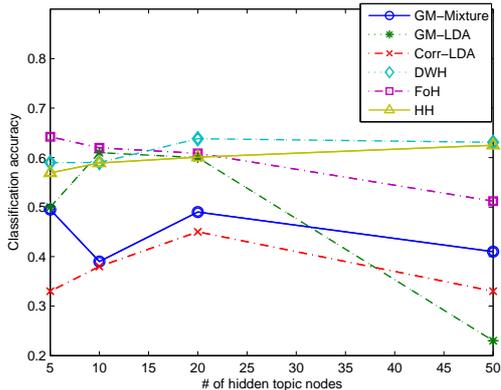


Figure 7: Classification performance of different models

representation of some kind for the video data. First, we implemented the approach used in [17], which learns a dual-wing harmonium (DWH) from the data and then builds a SVM classifier based on the latent semantic representations generated by DWH. This method is different from our approach in that it uses harmonium model only for data representation and leaves the classification task to a discriminative classifier, while our approach integrates the representation and classification. We also implemented three directed graphical models for representing video data, which are Gaussian multinomial mixture model (GM-Mixture), Gaussian multinomial latent Dirichlet allocation (GM-LDA), and correspondence latent Dirichlet allocation (Corr-LDA). The details of these models can be found in [1]. Similar to DWH, all the three directed models are used only for data representation, and each of them requires a SVM classifier for classification. To make the experiments with various models, learning algorithms, and numbers of latent topics tractable, we restrict this part of experiments to the 5 largest categories containing totally 1078 shots as *airplane*, *basketball*, *baseball*, *hockey*, and *weather news*.

Figure 7 shows the classification accuracies of the proposed FoH and HH models as well as the comparison methods including DWH, GM-Mixture, GM-LDA, and Corr-LDA. To be fair, all the models are implemented using the mean field variational method (MF) for learning and inference, except GM-Mixture which is implemented using expectation-maximization (EM) method. All the approaches are evaluated with the number of latent semantic topics set to 5, 10, 20, and 50, in order to study the relationship between performance and model complexity.

Several interesting observations can be drawn from Figure 7. First, the three undirected models as FoH,

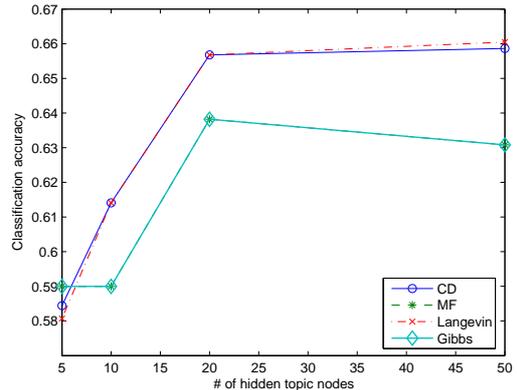


Figure 8: Classification performance of different approximate inference methods in hierarchical harmonium

HH, and DWH achieve significantly higher performance than the directed models as GM-Mixture, GM-LDA, and Corr-LDA, showing that the harmonium model is an effective tool for video representation and classification. Among them, FoH is the best performer at 5 and 10 latent semantic nodes, while DWH is the best performer at 20 and 50 latent nodes with HH as the close runner-up. Second, we find that the performance of FoH and HH is overall at the same level of DWH. Given that DWH uses a SVM classifier, this result is encouraging as it shows that our approach is comparable to the performance of a state-of-the-art discriminative classifier. On the other hand, our approach enjoys many advantages that SVM does not have. For example, FoH can be easily extended to accommodate a new category without re-training the whole model. Third, the performance of DWH and HH improves as the number of latent topics increases, which is intuitive because using more latent topics leads to better representation of the data. However, this trend is reversed in the case of FoH, which performs much better when using smaller number of latent topics. While a theoretical explanation of this is still unclear, in practice this is a good property of FoH since it achieves high performance with simpler models. Fourth, 20 seems to be a reasonable number of latent semantic topics for this data set, since further increasing the number of topics does not result in a considerable improvement of the performance.

Figure 8 shows the classification accuracies of HH model implemented using different approximate inference methods. From the graph, we can see that the Langevin and contrastive divergence (CD) methods have about the same performance, which is slightly higher than the performance of mean-field (MF) and Gibbs sampling. We also study the efficiency of these inference methods by examining the time they need to

reach convergence in training. The result shows that mean field is the most efficient (approx. 2 min), followed by CD and Langevin (approx. 9 min), and the slowest one is Gibbs sampling (approx. 49min). Therefore, Langevin and CD are good choices for the learning and inference of our models in terms of both efficiency and classification performance.

6 Conclusion

We have described two bipartite undirected models for semantic representation and classification of video data. The two models derive latent semantic representation of video data by jointly modeling the textual and image features of the data, and perform classification based on such latent representations. Experiments on TRECVID data have demonstrated that our models achieve satisfactory performance on video classification and provide insights to the internal structure and relationships of video categories. Approximate inference methods for the learning and application of our models have been discussed and compared.

Our hierarchical harmonium by nature does not restrict the number of categories an instance (shot) belongs to, since $P(Y_t = 1|\mathbf{x}, \mathbf{z})$ can be high for multiple Y_t . Therefore, an interesting future work is to evaluate the model with a multi-label data set, where each instance can belong to any number of categories. In this case, our method is actually a multi-task learning (MTL) method, and should be compared with other MTL approaches. Our models can also be improved using better low-level features as input. The region-based color histogram features are quite sensitive to scale and illumination variations. Features such as local keypoint features are more robust and can be easily integrated into our models. It is interesting to compare the latent semantic interpretations and classification performance using different features.

References

- [1] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual Int'l ACM SIGIR Conf. on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM Press.
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. volume 3, pages 993–1022, Cambridge, MA, USA, 2003. MIT Press.
- [3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- [4] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*.
- [5] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.
- [6] G. Iyengar and H. J. Nock. Discriminative model fusion for semantic concept detection and annotation in video. In *Proc. of the 11th ACM Int'l Conf. on Multimedia*, pages 255–258, New York, NY, USA, 2003. ACM Press.
- [7] M. I. Jordan. *Learning in Graphical Models: Foundations of Neural Computation*. The MIT press, 1998.
- [8] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.
- [9] I. Murray and Z. Ghahramani. Bayesian learning in undirected graphical models: Approximate mcmc algorithms. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-04)*, pages 392–399. AUAI press, 2004.
- [10] Y. Rui, R. Jain, N. D. Georganas, H. Zhang, K. Nahrstedt, J. Smith, and M. Kankanhalli. What is the state of our community? In *Proc. of the 13th annual ACM Int'l Conf. on Multimedia*, pages 666–668, New York, NY, USA, 2005. ACM Press.
- [11] A. Smeaton and P. Over. Trecvid: Benchmarking the effectiveness of infomration retrieval tasks on digital video. In *Proc. of the Intl. Conf. on Image and Video Retrieval*, 2003.
- [12] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis, 2005.
- [13] M. Welling and G. E. Hinton. A new learning algorithm for mean field boltzmann machines. In *ICANN '02: Proceedings of the Int'l Conf. on Artificial Neural Networks*, pages 351–357, London, UK, 2002. Springer-Verlag.
- [14] M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17*, pages 1481–1488, Cambridge, MA, 2004. MIT Press.
- [15] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proc. of the 12th annual ACM Int'l Conf. on Multimedia*, pages 572–579, 2004.
- [16] E. Xing, M. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence (UAI2003)*. Morgan Kaufmann Publishers, 2003.
- [17] E. Xing, R. Yan, and A. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *Proceedings of the 21th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-05)*. AUAI press, 2005.
- [18] R. Yan, J. Yang, and A. G. Hauptmann. Learning

query-class dependent weights in automatic video retrieval. In *Proc. of the 12th ACM Int'l Conf. on Multimedia*, pages 548–555. ACM Press, 2004.

APPENDIX

This is to show the derivation of the harmonium random fields (joint distribution) in the family-of-harmonium model. We start by introducing the *general form* of exponential-family harmonium [14] that has \mathbf{H} as the latent topic variables and \mathbf{X} and \mathbf{Z} as two types of observed data variables. This harmonium random field has the exponential form as:

$$p(\mathbf{x}, \mathbf{z}, \mathbf{h}) \propto \exp \left(\sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{jb} \eta_{jb} g_{jb}(z_j) + \sum_{kc} \lambda_{kc} e_{kc}(h_k) \right. \\ \left. + \sum_{ikac} W_{ia}^{kc} f_{ia}(x_i) e_{kc}(h_k) + \sum_{jkb} U_{jb}^{kc} g_{jb}(z_j) e_{kc}(h_k) \right).$$

where $\{f_{ia}(\cdot)\}$, $\{g_{jb}(\cdot)\}$, and $\{e_{kc}(\cdot)\}$ denote the sufficient statistics (features) of variables x_i , z_j , and h_k , respectively.

The marginal distributions, say, $p(\mathbf{x}, \mathbf{z})$, is then obtained by integrating out variables \mathbf{h} :

$$p(\mathbf{x}, \mathbf{z}) = \int_{\mathbf{h}} p(\mathbf{x}, \mathbf{z}, \mathbf{h}) d\mathbf{h} \\ \propto \exp \left(\sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{jb} \eta_{jb} g_{jb}(z_j) \right) \prod_k \int_{h_k} \exp \left(\sum_c \lambda_{kc} + \sum_{ia} W_{ia}^{kc} f_{ia}(x_i) + \sum_{jb} U_{jb}^{kc} g_{jb}(z_j) \right) e_{kc}(h_k) dh_k \\ = \exp \left(\sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{jb} \eta_{jb} g_{jb}(z_j) + \sum_k C_k(\{\hat{\lambda}_{kc}\}) \right)$$

and similarly we can derive:

$$p(\mathbf{x}, \mathbf{h}) \propto \exp \left(\sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{kc} \lambda_{kc} g_{kc}(h_k) + \sum_j B_j(\{\hat{\eta}_{jb}\}) \right) \\ p(\mathbf{z}, \mathbf{h}) \propto \exp \left(\sum_{jb} \eta_{jb} g_{jb}(z_j) + \sum_{kc} \lambda_{kc} e_{kc}(h_k) + \sum_i A_i(\{\hat{\theta}_{ia}\}) \right)$$

where the shifted parameters $\hat{\theta}_{ia}$, $\hat{\eta}_{jb}$ and $\hat{\lambda}_{kc}$ are defined as:

$$\hat{\theta}_{ia} = \theta_{ia} + \sum_{kc} W_{ia}^{kc} e_{kc}(h_k), \hat{\eta}_{jb} = \eta_{jb} + \sum_{kc} U_{jb}^{kc} e_{kc}(h_k) \\ \hat{\lambda}_{kc} = \lambda_{kc} + \sum_{ia} W_{ia}^{kc} f_{ia}(x_i) + \sum_{jb} U_{jb}^{kc} g_{jb}(z_j)$$

The functions $A_i(\cdot)$, $B_j(\cdot)$, and $C_k(\cdot)$ are defined as:

$$A_i(\{\hat{\theta}_{ia}\}) = \int_{x_i} \exp\left\{\sum_a \hat{\theta}_{ia} f_{ia}(x_i)\right\} dx_i \\ B_j(\{\hat{\eta}_{jb}\}) = \int_{z_j} \exp\left\{\sum_b \hat{\eta}_{jb} g_{jb}(z_j)\right\} dz_j \\ C_k(\{\hat{\lambda}_{kc}\}) = \int_{h_k} \exp\left\{\sum_c \hat{\lambda}_{kc} e_{kc}(h_k)\right\} dh_k$$

Further integrating out variables from these distribution give the marginal distribution of \mathbf{x} , \mathbf{z} , and \mathbf{h} .

$$p(\mathbf{x}) \propto \exp \left(\sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_j B_j(\{\hat{\eta}_{jb}\}) + \sum_k C_k(\{\hat{\lambda}_{kc}\}) \right)$$

$$p(\mathbf{z}) \propto \exp \left(\sum_{jb} \eta_{jb} g_{jb}(z_j) + \sum_i A_i(\{\hat{\theta}_{ia}\}) + \sum_k C_k(\{\hat{\lambda}_{kc}\}) \right)$$

$$p(\mathbf{h}) \propto \exp \left(\sum_{kc} \lambda_{kc} e_{kc}(h_k) + \sum_i A_i(\{\hat{\theta}_{ia}\}) + \sum_j B_j(\{\hat{\eta}_{jb}\}) \right)$$

We all the above marginal distributions, we are ready to derive the conditional distributions as:

$$p(\mathbf{x}|\mathbf{h}) = \frac{p(\mathbf{x}, \mathbf{h})}{p(\mathbf{h})} \propto \prod_i \exp \left(\sum_a \hat{\theta}_{ia} f_{ia}(x_i) - A_i(\{\hat{\theta}_{ia}\}) \right)$$

$$p(\mathbf{z}|\mathbf{h}) = \frac{p(\mathbf{z}, \mathbf{h})}{p(\mathbf{h})} \propto \prod_j \exp \left(\sum_b \hat{\eta}_{jb} g_{jb}(z_j) - B_j(\{\hat{\eta}_{jb}\}) \right)$$

$$p(\mathbf{h}|\mathbf{x}, \mathbf{z}) = \frac{p(\mathbf{x}, \mathbf{z}, \mathbf{h})}{p(\mathbf{x}, \mathbf{z})} \propto \prod_k \exp \left(\sum_c \hat{\lambda}_{kc} e_{kc}(h_k) - C_k(\{\hat{\lambda}_{kc}\}) \right)$$

The specific conditional distribution of \mathbf{x} , \mathbf{z} , and \mathbf{h} defined in Eq.(3.2), (3.3), and (3.4) are all exponential distributions. They can be mapped to the general forms above if we make the following definitions:

$$f_{i1}(x_i) = x_i$$

$$\theta_{i1} = \alpha_i, \hat{\theta}_{i1} = \alpha_i + \sum_k W_{ik} h_k$$

$$g_{j1}(z_j) = z_j, g_{j2}(z_j) = z_j^2$$

$$\eta_{j1} = \beta_j, \eta_{j2} = -1/2, \hat{\eta}_{j1} = \beta_j + \sum_k U_{jk} h_k$$

$$e_{k1} = h_k, e_{k2} = h_k^2$$

$$\lambda_{k1} = 0, \lambda_{k2} = -1/2, \hat{\lambda}_{k1} = \sum_i W_{ik} h_k + \sum_j U_{jk} h_k$$

Therefore, by plugging these definitions into general form of harmonium random field at the beginning of this appendix, we have the specific random field as:

$$p(\mathbf{x}, \mathbf{z}, \mathbf{h}) \propto \exp \left(\sum_i \alpha_i x_i + \sum_j \beta_j z_j - \sum_j \frac{z_j^2}{2} \right. \\ \left. - \sum_k \frac{h_k^2}{2} + \sum_{ik} W_{ik} x_i h_k + \sum_{jk} U_{jk} z_j h_k \right)$$

which is exactly the same as Eq.(3.5) except the latter one is defined for a specific category.