Harmonium Models for Video Classification

Jun Yang Rong Yan* Yan Liu* Eric P. Xing[1]

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
{juny, epxing}@cs.cmu.edu

*IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{yanr,liuya}@us.ibm.com

[1]To whom correspondence should be addressed.

Abstract

Accurate and efficient video classification demands the fusion of multi-modal information and the use of intermediate representations. Combining the two ideas into the one framework, we propose a series of probabilistic models for video representation and classification using intermediate semantic representations derived from multi-modal features of video. Based on a class of bipartite undirected graphical models named harmonium, we propose dual-wing harmonium (DWH) model that represents video shots as latent semantic topics derived by jointly modeling the transcript keywords and color-histogram features of the data. Our family-of-harmonium (FoH) and hierarchical harmonium (HH) model extends DWH by introducing variables representing category labels of data, which allows data representation and classification to be performed in the same model. Our models are among the few attempts of using undirected graphical models for representing and classifying video data. Experiments on a benchmark video collection show different semantic interpretations of video data under our models, as well as superior classification performance in comparison with several directed models.

1 Introduction

Classifying video data into semantic categories, sometimes known as semantic video concept detection, is an important research topic. This task is challenging because video data contain multiple data types including video frames as images, transcript text, speech, audio, each bearing correlated and complementary information essential to conveying data semantics. The fusion of such multi-modal information is regarded as a key research problem [11], and has been a widely used technique in video classification and retrieval methods. Many fusion strategies have been proposed, varying from early fusion [14], 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 [14, 7, 20, 17]. Empirical results show that the methods based on the fusion of multi-modal information outperforms those based on any single type of information in both classification and retrieval tasks.

Another trend in video classification is the search of low-dimensional, intermediate representations of video data in order to replace the high-dimensional raw features such as color histograms and term vectors. The reason is that intermediate representations would make sophisticated classifiers such as support vector machines [3] computationally efficient, which would be more expensive when applied on raw features. Moreover, using intermediate representations holds the promise of better interpretation of data semantics, and may lead to superior classification performance. Related work along this direction ranges from conventional dimension-reduction methods such as principal component analysis (PCA) and Fisher linear discriminant (FLD) [5], to the more recent probabilistic "topic models" such as probabilistic latent semantic indexing (pLSI) [6], latent Dirichlet allocation (LDA) [2], exponential-family harmonium (EFH) [16]. While most of these models are initially developed only for single-modal data such as textual documents, extensions of them [1] have been studied recently in order to model data with multiple types of inputs (a.k.a multi-modal data) such as captioned images and video.

The key insights for video classification from previous works appear to be combining multi-modal information and using intermediate representations. The goal of this paper is to propose a series of probabilistic models for representing and classifying video data by taking advantages of both insights. Our models are based on a class of bipartite, undirected graphical models (i.e., random fields) called harmoniums [16]. The first model, *dual-wing harmonium* (DWH), derives intermediate representation as a set of *latent semantic topics* of video shots, by jointly modeling the correlated information in the image regions and transcript keywords associated with the video shots. The derived latent topics are then used as semantic features for classifying video shots. The other two models, *family-of-harmonium* (FoH) and *hierarchical harmonium* (HH), extend DWH by explicitly incorporating the category label(s) of data into the model, which allows the classification and representation to be accomplished in a unified framework. Specifically, FoH consists of a set of category-specific DWH models, each modeling the video data from one specific category, and it assigns a video shot to the category with the highest probability. In contrast, HH introduces category labels as another layer of hidden variables into a DWH model, and performs classification through the inference of these label variables.

The proposed models differ from existing models for text/multimedia data in several important aspects. First, our models are among the first few undirected topic models for bi-modal or multi-modal data such as video, as most of the existing models are directed and they are mainly proposed for single-modal data such as text documents. Besides providing an important alternative for modeling video data, our models do offer unique properties not supported by their directed counterparts, among which is fast inference due to the conditional independence between latent variables. Furthermore, two of our models, FoH and HH, incorporate category labels as (hidden) model variables, which allows us to classify unlabeled data by directly inferencing the distribution of the label variables. In comparison, most existing models [2, 1, 6, 16] can be only used for deriving intermediate data representation as latent semantic topics, and one has to build separate classifiers on top of the derived representation if

classification is to be performed. Therefore, another advantage of our approach lies in the unification of representation and classification in the same model, which avoids the overhead of building separate classifiers. More importantly, by considering the interactions between latent semantic topics and category labels, our models may be able to learn better intermediate representations so as to reflect the category information from 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, 6, 16].

The notations used in the paper follow the convention of probabilistic graphical models. Uppercase characters represent random variables, and 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 all the illustrations, shaded circles represent observed nodes while empty circles represent hidden (latent) nodes. Each node in a graphical model is associated with a random variable, so we use the term node and variable interchangeably in this paper.

In Section 2 we review the related work on the fusion of multimodal video features as well as representation models for text and multimedia data. A brief introduction of harmoniums is presented in Section 3. We propose the three harmoniums models for video data in Section 4, and discuss their learning algorithms in Section 5. In Section 6, we show the experiment results and illustrate interesting interpretation of the data from TRECVID video collection. The conclusions and future work are discussed in Section 7.

2 Related Works

As pointed out in [11], the processing, indexing, and fusion of the 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 and achieve better performance than using any single type of feature. This idea has been widely used in many existing methods. The fusion strategies vary from early fusion [14], 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 [14, 7, 20, 17]. 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 [14]. The approach presented in this paper takes neither approach; instead, it derives the latent semantic representation of the video data by jointly modeling the multimodal low-level features, so that the fusion takes place somewhere between early fusion and late fusion.

There are many approaches to obtaining 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 widely-used alternatives for dimension reduction. Recently, there are also many studies on modeling the latent semantic topics of the text and multimedia data. For example, latent semantic indexing (LSI) by Deerwester et al. [4] transforms term counts linearly into a low-dimensional semantic eigenspace, and the idea was later extended by Hofmann to probabilistic LSI (pLSI) [6]. 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. [16] is bipartite undirected graphical model consisting a layer of latent nodes representing semantic aspects and a layer of observed nodes representing the raw features.

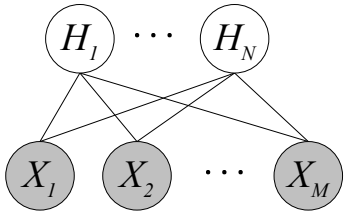


Figure 1: The basic harmonium model

In practice, the methods mentioned above are mainly used for transforming the high-dimensional raw features 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., SVMs) based on such latent semantic representations. In this paper, two of the proposed models, namely FoH and HH, provide a unified approach that integrates representation and classification in the same framework. They not only achieve satisfactory classification performance, but also provide interesting insights into the data semantics, such as the internal structure of each category and the relationships between different categories. Fei-Fei et al. [9] used a unified model for representing and classifying natural scene images by introducing category variables into the LDA model.

3 The Basic Harmonium Model

The harmoniums, which were originally studied by Smolensky [13] in his "harmonium theory", refer to a family of bipartite undirected graphical models (a.k.a random fields) that consist of two layers of nodes. Figure 1 shows a basic harmonium model, where nodes $\mathbf{X} = \{X_i\}$ at the bottom layer denote the *observed* data and the nodes $\mathbf{H} = \{H_i\}$ at the topic layer model the *latent* semantic topics of the data. Depending on the specific application, the data nodes $\mathbf{X}$ can represent keyword counts of a text document or image histogram features of an image, and the latent topic nodes $\mathbf{H}$ constitute a low-dimensional summarization of the data that capture the critical information in the raw data.

The bipartite topology of a harmonium ensures that the nodes within the same layer are conditionally independent given the nodes in the other layer. This makes possible a convenient constructive definition of the harmonium distribution based on two between-layer conditional distribution $p(\mathbf{x}|\mathbf{h})$ and $p(\mathbf{hx})$, both of which factorize over individual nodes as $p(\mathbf{x}|\mathbf{h}) = \prod_i p(x_i|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{x}) = \prod_j p(h_j|\mathbf{x})$. Welling et al. [16] proposed a special class of harmonium models called *exponential-family harmonium* (EFH), where these conditionals are from exponential family:

$$\begin{aligned} p(\mathbf{x}|\mathbf{h}) &= \prod_i p(x_i|\mathbf{h}) \propto \prod_i \exp\left\{(\theta_i + \sum_j W_{ij} g(h_j)) f(x_i)\right\} \\ p(\mathbf{h}|\mathbf{x}) &= \prod_j p(h_j|\mathbf{x}) \propto \prod_j \exp\left\{(\lambda_j + \sum_i W_{ij} f(x_i)) g(h_j)\right\} \end{aligned} \tag{3.1}$$

where $\{f(x_i)\}$ and $\{g(h_j)\}$ are the sufficient statistics (or features) of node x_i and h_j; $\{\theta_i\}$, $\{\lambda_j\}$, and $\{W_{ij}\}$ are model parameters, which can be learned from the data. The partition function (i.e., normalizer) in these distributions are not explicitly shown, and therefore we use a proportional sign instead of an equal size in Eq.(3.1). We see that the data nodes $\mathbf{x}$ and the topic nodes $\mathbf{h}$ are coupled through an interaction term W_{ij}, so that the values of the topic nodes $\mathbf{h}$ affect the distribution of $\mathbf{x}$, and vice versa. This ensures that the latent topic variables $\mathbf{h}$ are in "harmony" with the data variables $\mathbf{x}$, so that $\mathbf{h}$ preserve most of the information in $\mathbf{x}$.

Welling et al. [16] showed that these local conditionals precisely map to the following harmonium

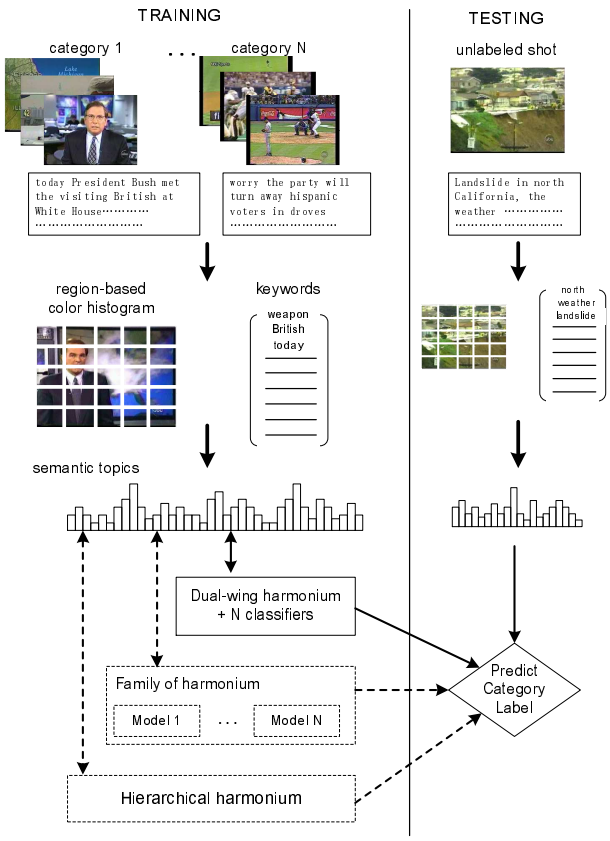


Figure 2: A sketch of our approach to video classification

random fields (joint distribution):

$$p(\mathbf{x}, \mathbf{h}) \propto \exp\left\{ \sum_i \theta_i f(x_i) + \sum_j \lambda_j g(h_j) + \sum_{ij} W_{ij} f(x_i) g(h_j) \right\}$$

After the parameters $\{\theta_i\}, \{\lambda_j\}, \{W_{ij}\}$ are learned, the harmonium model can be used to infer the latent topic nodes $\mathbf{h}$ from the observed data nodes $\mathbf{x}$. Due to the conditional independence between the latent nodes, the inference of $\mathbf{h}$ is very efficient. This is a nice property not provided by the directed graphical models, which typically do not have such conditional independence. On the other hand, however, there is no *marginal* independence for either data or latent topic nodes in a harmonium. Therefore, learning harmoniums is usually more difficult due to the presence of the global partition function.

The basic harmonium has been used to derive the latent semantic topics from the keyword features of text documents [16]. However, it is inadequate for modeling complex data such as video data that contain multiple types of inputs (features) following distributions of different families. We describe a series of extensions of the basic harmonium for modeling and classifying video data.

4 Harmonium Models for Video Data

A sketch of our approach to video classification is illustrated in Figure 2. We classify video data in the form of video shots, which are short video segments with length varying from a few seconds to half minute or even longer. As video contain both textual and imagery data, we represent each video shot as a bag of keywords (extracted from the video closed-captions or via speech recognition systems), and a set of fixed-sized image regions (extracted from a representative frame or keyframe of the video shot).

Each region is described by its color histogram feature. In the training phase, the goal is to build a model of a certain type that derives the latent semantic topics of video data and captures the latent topics (and their combinations) that best describe each category. During the testing phase, we extract the keywords and color features from an unlabeled video shot, and then use them as features to predict which category this shot belongs to. The three proposed models, DWH, FoH, and HH, differ in the way the data are represented and classified.

4.1 Notations and definitions The random variables and parameters in our harmonium models are defined 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 $\mathbf{x} = (x_1, ..., x_N)$ 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 $\mathbf{z} = (z_1, ..., z_M)$ denotes color-histogram features of the keyframe in 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 $\mathbf{h} = (h_1, ..., h_K)$ 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 $y \in \{1, ..., T\}$ indicates the category this shot belongs to, where T is the total number of categories. In hierarchical harmonium, the labels are represented by a vector $\mathbf{y} = (y_1, ..., y_T)$, 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 three harmonium models presented below have different parameters. The parameters of a dual-wing harmonium are denoted as $\theta^y = (\alpha, \beta, W, U)$. A family-of-harmonium contains a set of category-specific dual-wing harmoniums, and each one has parameters as $\theta^y = (\pi_y, \alpha^y, \beta^y, W^y, U^y)$ where y is the category label. A hierarchical harmonium has a single set of parameters as $\theta = (\alpha, \beta, \tau, W, U, V)$.

4.2 Dual-wing harmonium (DWH) As illustrated in Figure 3, *dual-wing harmonium* (DWH) extends the basic harmonium by introducing two "wings" of data nodes in order to represent the bi-modal information of video data. The nodes in the top layer represent the *latent* semantic topics $\mathbf{H} = \{H_k\}$ of a video shot s; nodes in the bottom layer consists of two sets of *observed* variables: $\mathbf{X} = \{X_i\}$ representing the keyword feature of the video shot and $\mathbf{Z} = \{Z_j\}$ representing the the region-based color feature of the shot. Thus, DWH models the low-level (keyword and color) features of a video shot as well as its latent semantic topics as two types of representations that influence each other. 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 detailed below.

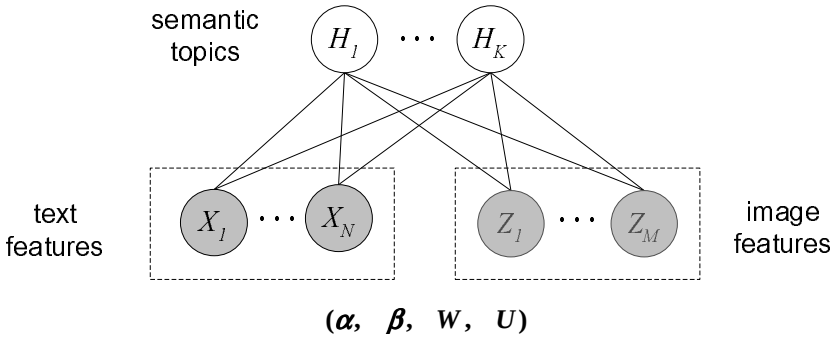


Figure 3: Dual-wing harmonium model

Text feature: The variable x_i indicating the presence/absence of term $i \in \{1, ..., N\}$ in the vocabulary follows a distribution as:

$$P(X_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(-\alpha_i - \sum_k W_{ik} h_k)} \tag{4.2}$$
$$P(X_i = 0|\mathbf{h}) = 1 - P(X_i = 1|\mathbf{h})$$

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

Image feature: The color-histogram feature z_j of the j^{th} region in the keyframe of the shot admits a conditional multivariate Gaussian distribution as:

$$p(z_j|\mathbf{h}) = \mathcal{N}(z_j|\Sigma_j(\beta_j + \sum_k U_{jk} h_k), \Sigma_j) \tag{4.3}$$

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

Latent semantic topics: Finally, each latent topic variable h_j follows a unit-variance Gaussian distribution whose mean is determined by a weighted combination of the keyword feature $\mathbf{x}$ and the color feature $\mathbf{z}$:

$$p(h_k|\mathbf{x}, \mathbf{z}, c) = \mathcal{N}(h_k| \sum_i W_{ik} x_i + \sum_j U_{jk} z_j, 1) \tag{4.4}$$

where W_{ik} and U_{jk} are the same parameters used in Eq.(4.2) and (4.3). Similarly, $p(\mathbf{h}|\mathbf{x}, \mathbf{z}) = \prod_k p(h_k|\mathbf{x}, \mathbf{z})$ 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:

$$p(\mathbf{x}, \mathbf{z}, \mathbf{h}) \propto \exp\left\{ \sum_i \alpha_i x_i + \sum_j \beta_j z_j - \sum_j \frac{z_j^2}{2} - \sum_k \frac{h_k^2}{2} + \sum_{ik} W_{ik} x_i h_k + \sum_{jk} U_{jk} z_j h_k \right\} \tag{4.5}$$

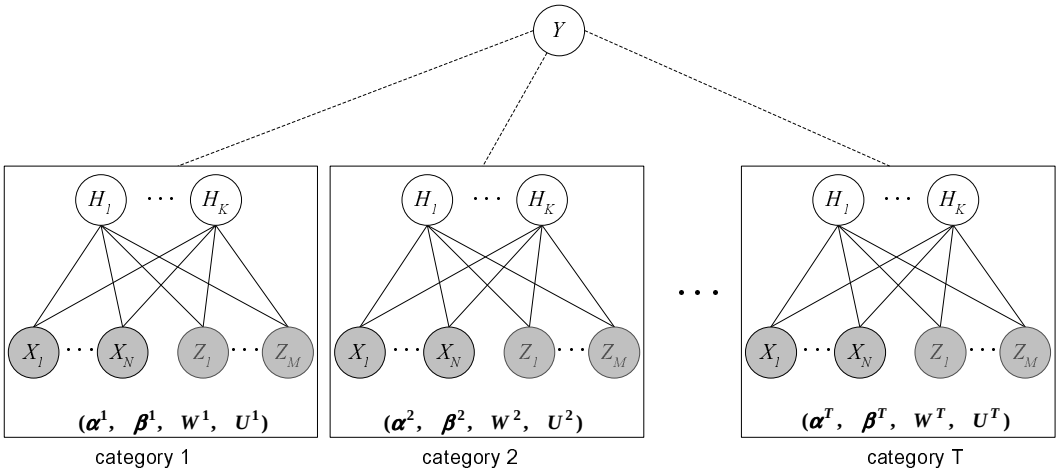


Figure 4: The family-of-harmonium model

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

By integrating out the hidden variables $\mathbf{h}$ in Eq.(4.5), we obtain the marginal distribution over the observed keyword and color features of a video shot:

$$p(\mathbf{x}, \mathbf{z}) \propto \exp\left\{ \sum_i \alpha_i x_i + \sum_j \beta_j z_j - \sum_j \frac{z_j^2}{2} + \frac{1}{2} \sum_k (\sum_i W_{ik} x_i + \sum_j U_{jk} z_j)^2 \right\} \tag{4.6}$$

which also contains a hidden partition function in this distribution.

The parameters of a DWH model, $\theta = (\alpha, \beta, W, U)$, is learned by maximizing the likelihood of a set of video shots, where the likelihood function is defined by Eq.(4.6). Due to the presence of the global partition function, the learning process requires approximate inference methods, which will be discussed in Section 5. Note that in Eq.(4.2) we define the variance of the latent variables given the input variables to one in order to simplify the parameter estimation. Introducing a covariance matrix Σ can offer additional freedom for joint distribution $p(\mathbf{x}, \mathbf{z}, \mathbf{h})$, but it would not lead to more general representations in terms of probability $p(\mathbf{x}, \mathbf{z})$ [16].

It is important to note that DWH is a model proposed only for representing video data: given the text and color features $\mathbf{x}$ and $\mathbf{z}$ of a video shot, one can infer the latent semantic topics $\mathbf{h}$ of the video shot using a DWH model. The model by itself cannot be used directly for classification, because it contains no variables representing the category labels of data. To do classification, one needs to first represent video data by their latent semantic topics (i.e., treating latent variables $\mathbf{h}$ as a feature vector), and build classifiers using classification models such as Support Vector Machines (SVMs) based on such latent semantic representation. The learning of the DWH model is "unsupervised" in the sense that it does not involve data labels. In practice, we build only a *single* DWH model from all the available video shots despite the number of categories they belong to. Moreover, the DWH model does not have to be updated or re-trained when new categories (i.e., unseen in the training data) arrive; one can simply build more classifiers for these new categories on the representations derived from the same DWH model.

4.3 Family-of-harmonium (FoH) The Family-of-harmonium (FoH) model extends the DWH model in order integrate classification and representation in the same model. The FoH model uses the DWH

model as its basic building block. As illustrated in Figure 4, a FoH model consists of a set of T category-specific harmoniums, where each harmonium is a DWH that models video shots from a specific category. The number of component harmoniums is equal to the number of categories. On top of these component harmoniums, a node $Y \in \{1, ..., T\}$ representing the category label is introduced as a "switch variable" to indicate the specific harmonium used for modeling a given video shot. The semantics of a FoH model is apparent from its structure: given the category of a video shot, it uses the harmonium corresponding to that category to model that video shot.

All the component harmoniums in FoH share exactly the same structure, because they are all DWH models with the same number of input and latent nodes and same forms of distributions, except that each harmonium owns a unique set of parameters $(\alpha^y, \beta^y, W^y, U^y)$ indexed by the category label y. The label variable Y is only an indicator of the specific harmonium used for modeling the video shot. Therefore, in Figure 4 Y is not actually linked to any nodes in the component harmoniums, and it only appears as the subscript of model parameters in the distribution function to be presented below.

The distribution of a FoH model can be easily constructed from the distribution of each component DWH. For each DWH, the conditionals of variables of each type, namely $\mathbf{x}$, $\mathbf{z}$, and $\mathbf{h}$ follow the distribution defined in Eq.(4.2), (4.3), and (4.4), except that the parameters are indexed by the category label y. Therefore, the likelihood function of a component DWH given the category, $p(\mathbf{x}, \mathbf{z}|y)$, would have exactly the same form as the joint distribution of DWH defined in Eq.(4.6). The category label Y, the only new variable in FoH, follows a prior multinomial distribution as:

$$p(y) = \text{Multi}\ (\pi_1, ..., \pi_T), \tag{4.7}$$

where $\sum_{t=1}^{T} \pi_t = 1$. The marginal distribution (likelihood) of a labeled video shot in a FoH model can be decomposed into a category-specific marginal and a prior over the categories:

$$p(\mathbf{x}, \mathbf{z}, y) = p(\mathbf{x}, \mathbf{z}|y)p(y) \propto \pi_y \exp\left\{ \sum_i \alpha_i^y x_i + \sum_j \beta_j^y z_j - \sum_j \frac{z_j^2}{2} + \frac{1}{2} \sum_k (\sum_i W_{ik}^y x_i + \sum_j U_{jk}^y z_j)^2 \right\} \tag{4.8}$$

Learning a FoH model is equivalent to learning T independent DWH models, where each component DWH is learned from video shots from the corresponding category by maximizing the likelihood function defined in Eq.(4.8). Therefore, the learning method used for DWH is readily applicable to the learning of FoH, which will be discussed in Section 5.

For classification, FoH behaves like a maximum likelihood classifier. That is, it examines the probability function $p(\mathbf{x}, \mathbf{z}, y)$ of a video shot under each of the component harmoniums, and assigns the shot to the category corresponding to the harmonium with the highest probability. This is because given Baye's rule the posterior probability of category label is proportional to the data likelihood as $p(y|\mathbf{x}, \mathbf{z}) \propto p(\mathbf{x}, \mathbf{z}|y)p(y)$. We can further simplify this by assuming that the category prior is a uniform distribution, e.g., $p(y) = 1/T$. Thus, we can predict the category of a shot by comparing its class conditional $p(\mathbf{x}, \mathbf{z}|y)$ under each harmonium y, which can be computed from Eq.(4.8). The harmonium that "best fits" the shot determines its category.

4.4 Hierarchical harmonium (HH) Hierarchical harmonium (HH) extends the basic DWH model in a different way in order to make it directly applicable to classification tasks. Instead of building a separate harmonium for each category, it introduces category labels as another (the third) layer of nodes into a single dual-wing harmonium, making it an undirected model of three layers. In Figure 5, the label variables $\mathbf{Y} = \{Y_1, ..., Y_T\}$ with $Y_t \in \{0, 1\}$ indicate a shot's membership with each category, and they form a bipartite subgraph with the latent topic variables $\mathbf{H}$ on top of the bipartite subgraph between $\mathbf{H}$ and the input $\mathbf{X}$ and $\mathbf{Z}$. Unlike a FoH model, the label variables $\mathbf{Y}$ in HH are linked to the other

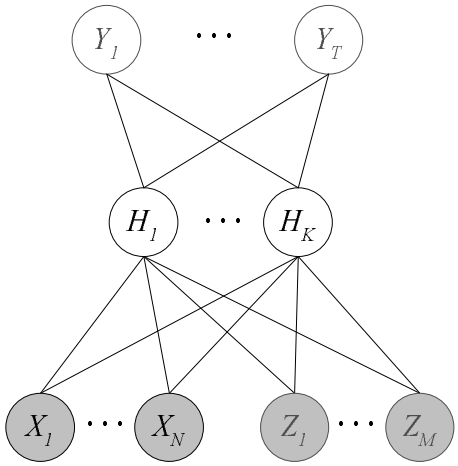


Figure 5: Hierarchical harmonium model

variables in the model. In fact, there is a link between any Y_t and H_j but not between two Y_t, so the conditional independence property of harmoniums is preserved. Another difference is that a HH model contains only a single harmonium while a FoH model contains a set of harmoniums for all categories.

In a HH model, the conditional distribution of $\mathbf{x}$ and $\mathbf{z}$ stay the same as those in the DWH model, which are defined by Eq.(4.2) and Eq.(4.3), respectively. Each label variable Y_i follows a Bernoulli distribution parameterized by the latent variables $\mathbf{h}$:

$$\begin{aligned} P(Y_t = 1|\mathbf{h}) &= \frac{1}{1 + \exp(-\tau_t - \sum_k V_{tk} h_k)} \\ P(Y_t = 0|\mathbf{h}) &= 1 - P(Y_t = 1|\mathbf{h}) \end{aligned} \tag{4.9}$$

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 distribution of the class label in logistic regression [5], 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 variable h_k needs to be modified to incorporate the interactions between label variables $\mathbf{y}$ and the topic variables $\mathbf{h}$:

$$p(h_k|\mathbf{x}, \mathbf{z}, \mathbf{y}) = \mathcal{N}(h_k| \sum_i W_{ik} x_i + \sum_j U_{jk} z_j + \sum_t V_{tk} y_t, 1) \tag{4.10}$$

This shows that the distribution of the latent semantic topics $\mathbf{h}$ are not only affected by data features $\mathbf{x}$ and $\mathbf{z}$, but also affected by their labels $\mathbf{y}$. This is significantly different from the DWH and FoH model, as well as directed graphical models such as LDA [2], where the distribution of latent variables only depend on the input features. In this sense, the latent semantic topics derived from a HH model are "supervised" while those derived by other models are "unsupervised".

With the incorporation of label variables, the joint distribution (i.e., random field) of a HH model becomes:

$$p(\mathbf{x}, \mathbf{z}, \mathbf{h}, \mathbf{y}) \propto \exp\left\{ \sum_i \alpha_i x_i + \sum_j \beta_j z_j - \sum_j \frac{z_j^2}{2} + \sum_t \tau_t y_t - \sum_k \frac{h_k^2}{2} + \sum_{ik} W_{ik} x_i h_k + \sum_{jk} U_{jk} z_j h_k + \sum_{tk} V_{tk} y_t h_k \right\} \tag{4.11}$$

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

$$p(\mathbf{x}, \mathbf{z}, \mathbf{y}) \propto \exp\left\{ \sum_i \alpha_i x_i + \sum_j \beta_j z_j - \sum_k \frac{z_k^2}{2} + \sum_t \tau_t y_t + \frac{1}{2} \sum_k (\sum_i W_{ik} x_i + \sum_j U_{jk} z_j + \sum_t y_t V_{tk})^2 \right\} \tag{4.12}$$

The parameters of the HH model, $\theta = (\alpha, \beta, \tau, W, U, V)$, are estimated by maximizing the likelihood function defined by Eq.(4.12). Despite the introduction of label variables $\mathbf{Y}$, the learning procedure for HH is in spirit similar to that for DWH and FoH, and can be easily extended from the latter.

The classification is performed in a way different from either in DWH or in FoH. Since data labels are represented as model variables $\mathbf{Y}$, we can predict the category of an unlabeled video shot by inferring the label variables $\mathbf{Y}$ from its text and image features. This is done by computing the conditional probability $p(Y_t = 1|\mathbf{x}, \mathbf{z})$ for each label variable Y_t. We can assign the shot to the category with the highest probability, i.e., $t^* = \text{argmax}_t p(Y_t = 1|\mathbf{x}, \mathbf{z})$. If a video shot may belong to more than one category, we can assign it to categories with probability above a given threshold. There is, however, no analytical solution to inferring $p(Y_t = 1|\mathbf{x}, \mathbf{z})$. Various approximate inference methods are available to solve this problem, as further discussed in Section 5.

4.5 Discussion There are several interesting differences and connections between the three harmonium models we have proposed.

- First of all, DWH is a representation model while FoH and HH are classification models. DWH has no variables representing category labels, can be trained without data labels, and cannot be used directly for classification. It only derives the latent semantic representations of data, and one needs to build separate classifiers to classify the data based on such representations. In comparison, FoH and HH incorporate label variables and need to be trained using labeled data. They not only derive the latent representation of the data but also perform classification within the same model. It is difficult to say theoretically which model is better because they are for different purposes. But if classification is the only purpose, FoH and HH does provide a more integrated and efficient approach by avoiding the need of training separate classifiers.

- Furthermore, the meanings of the derived latent topics are different in these models. In DWH and HH, the latent topics represent the "common topics" of the data, since all the data share the same set of latent topics. The latent topics in HH are likely to be different from those in DWH, because they are "supervised" by the category labels and presumably contain more discriminative information. The latent topics in HH also help to reveal the connections between various categories. In FoH, since each component harmonium is built for a specific category, the latent topics in each harmonium capture the internal structure of that category, i.e., they represent the themes or subcategories in that particular category. There are no correspondences between the latent topics across different harmoniums: the first topic in one harmonium is unrelated to the first topic in another.

- The three models also differ in terms of efficiency and flexibility. On a fixed collection, training a HH model is less expensive than training a FoH model as the latter involves training multiple harmoniums. Training a DWH model is not expensive by itself, but it can be costly to train separate classifiers on top it depending on the classification algorithm used. When it comes to incorporating a new category, FoH can accommodate the new category by adding another harmonium trained from its data without changing the harmoniums for existing categories, and DWH does not even have to be updated except that a new classifier needs to be built for the new category. Introducing

a new category is more expensive in a HH model because that means adding another label node into the model, which requires re-training of the whole model as its structure has been changed.

To provide deeper insights of our models, we also compare them to other topics models for text and multimedia data, including PCA, pLSI [6], LDA [2] and its variants GM-LDA and Corr-LDA [1], exponential-family harmonium (EFH) [16, 19]. Our models are among the first few attempts to use undirected topic models, since most existing topic models are directed. Although there is no conclusion yet as to which one is better, our models offer appealing properties such as easy inference due to the conditional independence. Also note that the majority of topic models are for single-modal data, usually text documents, while our models join GM-LDA and Corr-LDA to be the few topic models for bi-modal data such as video and captioned images. Finally, our FoH and HH model integrates representation and classification in the same framework. In contrast, most existing topic models are only intended for data representation. The Bayesian hierarchical model for scene classification proposed by Fei-Fei et al. [9], which is extended from LDA, is a counterpart of FoH in the directed models.

5 Learning and inference

The parameters of the three harmonium models in Section 4 are learned under the maximum likelihood principle using gradient ascent and approximate inference techniques. In this section, we use the learning of HH as an example to describe the general procedure of learning a harmonium model, because HH is structurally the most complex model among the three. The learning algorithm for DWH can be "reduced" from the algorithm for HH, and learning a FoH is equivalent to learning multiple DWH models for different categories.

Given a labeled set of video shots $\mathcal{X} = \{\mathbf{x}_n, \mathbf{z}_n, y_n\}_{n=1}^N$, the parameters of a HH model $\theta = (\alpha, \beta, \tau, W, U, V)$ is estimated by maximizing the log-likelihood of the data defined by Eq.(4.12). Due to the complexity of the model, there is no closed-form solution to the maximization problem and we have to resort to an iterative method like gradient ascent. The learning rules (i.e., the gradients) can be obtained by setting the derivatives of Eq.(4.12) with respect to model parameters:

$$\begin{aligned} \delta\alpha_i &= \langle x_i \rangle_{\tilde{p}} - \langle x_i \rangle_p, \quad \delta\beta_j = \langle z_j \rangle_{\tilde{p}} - \langle z_j \rangle_p, \quad \delta\tau_t = \langle y_t \rangle_{\tilde{p}} - \langle y_t \rangle_p \\ \delta W_{ik} &= \left\langle x_i h'_k \right\rangle_{\tilde{p}} - \left\langle x_i h'_k \right\rangle_p, \quad \delta U_{jk} = \left\langle z_j h'_k \right\rangle_{\tilde{p}} - \left\langle z_j h'_k \right\rangle_p, \quad \delta V_{tk} = \left\langle y_t h'_k \right\rangle_{\tilde{p}} - \left\langle y_t h'_k \right\rangle_p \end{aligned} \tag{5.13}$$

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 other undirected graphical models, there is a global normalizer (a.k.a partition function) in the likelihood function of harmonium Eq.(4.12), which makes directly computing $\langle \cdot \rangle_p$ intractable. Instead, we need approximate inference methods to estimate these model expectations $\langle \cdot \rangle_p$. We explored four approximate inference methods in our work, which are briefly discussed below. Besides the learning process, the inference of the conditional distribution of label variables $p(Y_t = 1 | \mathbf{x}, \mathbf{z})$ in a HH model is also intractable and requires using approximate inference methods.

5.1 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 [18]. We use the naive version of MF, where the joint probability p is approximated by an surrogate distribution q as a product of *singleton* marginals over the variables:

$$q(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{h}) = \prod_i q(x_i | \nu_i) \prod_j q(z_j | \mu_j, I) \prod_t q(y_t | \lambda_t) \prod_k q(h_k | \gamma_k) \tag{5.14}$$

where the singleton marginals are defined as $q(x_i) \sim$ Bernoulli (ν_i), $q(z_j) \sim N(\mu_j, I)$, $q(y_t) \sim$ Bernoulli (λ_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:

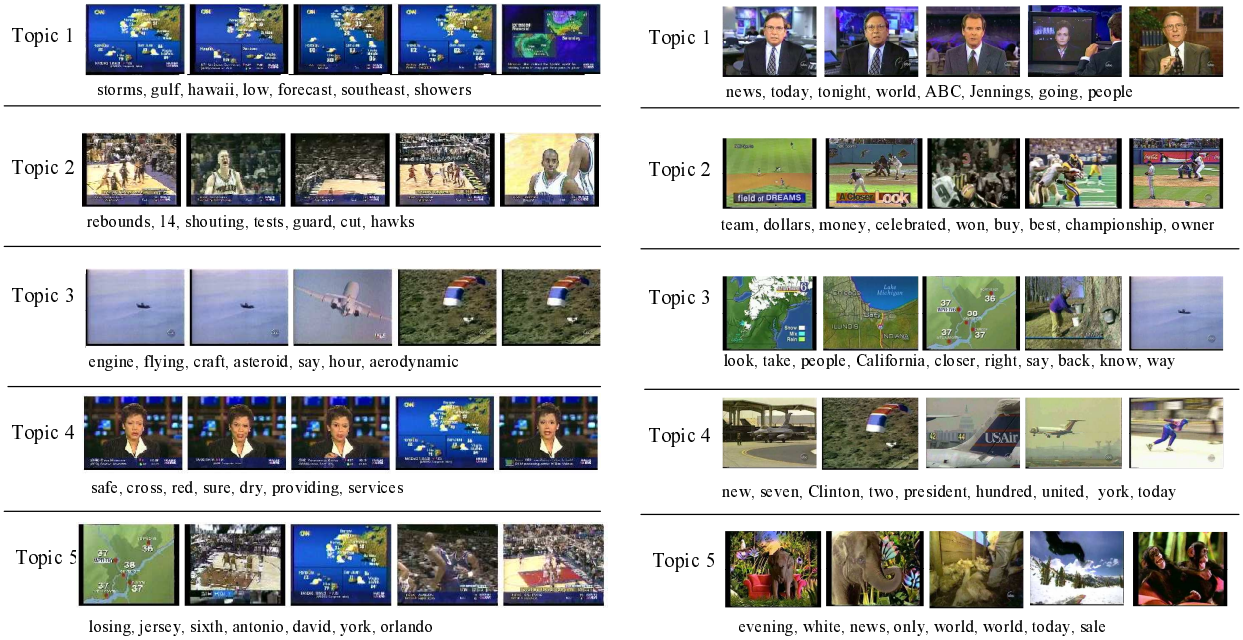
$$\nu_i = \sigma(\alpha_i + \sum_k W_{ik}\gamma_k)$$
$$\mu_j = \beta_j + \sum_k U_{jk}\gamma_k$$
$$\lambda_t = \sigma(\tau_t + \sum_k V_{tk}\gamma_k)$$
$$\gamma_k = \sum_i W_{ik}v_i + \sum_j U_{jk}\mu_j + \sum_t V_{tk}\lambda_t$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function. We iteratively update the variational parameters using the above fixed-point equations until they converge, and then the surrogate distribution q is fully specified. We replace the intractable model expectations $\langle\cdot\rangle_p$ with $\langle\cdot\rangle_q$ in Eq.(5.13), which are easy to compute from the fully factorized surrogate distribution q. Then, we can update the model parameters using the learning rules defined in Eq.(5.13).

It is important to note that, when using gradient ascent method with mean file, the learning procedure would contain two nested loops: the outer loop iteratively updates the model parameters using the learning rules Eq.(5.13), while the inner loop iteratively updates the variational parameters in order to approximate the model expectations in the learning rules. Whenever the model parameters are updated (and so are the model distribution p), the whole inner loop needs to be executed to recompute the surrogate distribution q to approximate the updated model distribution p. Using gradient ascent with other iterative approximate inference methods such as Gibbs sampling would also result in a learning procedure with nested loops.

5.2 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 [8]. 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 a HH model, we define the sampling order to be $y_1, \ldots, y_T, h_1, \ldots, h_K$, as the other variables are given as input. This means we first sample each y_t from the conditional distribution defined in Eq.(4.9) using the current values of $\{h_j\}$, and then sample each h_j according to Eq.(4.10) using the sampled values of $\{y_t\}$, and repeat this process iteratively. 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 samples collected *after* the burn-in period to approximate the true expectation $\langle\cdot\rangle_p$. The number of "burn-in" iterations and samples is at least thousands and typically around tens of thousands. Therefore, although this method guarantees accurate approximations, it is computationally intensive.

5.3 Contrastive divergence An alterative to exact gradient ascent search based on the learning rules in Eq.(5.13) is the contrastive divergence (CD) algorithm [15] proposed by Hinton and Welling that approximates the gradient learning rules. In each step of the gradient ascent, instead of computing the model expectation $\langle\cdot\rangle_p$, CD starts from the empirical values as the initial samples, runs the Gibbs sampling for up to only *a few* iterations and uses these limited samples to approximate the model expectation $\langle\cdot\rangle_p$. It has been proved that the final values of the parameters by this kind of updating will converge to the maximum likelihood estimation [15]. In our implementation, we compute $\langle\cdot\rangle_q$ from a large



(a) Topics of DWH (b) Topics of HH

Figure 6: An illustrated of 5 latent topics derived by (a) the DWH model and (b) the HH model from the data collection. Each topic is shown by the top 10 keywords and top 5 key-frames extracted from the most related video shots.

number of samples obtained by running only *one* step of Gibbs sampling with different initializations. Obviously, CD is substantially more efficient than the Gibbs sampling method since the "burn-in" process is skipped.

5.4 The uncorrected Langevin method The uncorrected Langevin method [10] is originated from the Langevin Monte Carlo method by accepting all the proposal moves. It makes use of the 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:

$$\lambda_{ij}^{\text{new}} = \lambda_{ij} + \frac{\epsilon^2}{2}\frac{\partial}{\partial \lambda_{ij}} \log p(X, \lambda) + \epsilon n_{ij} \tag{5.15}$$

where $n_{ij} \sim \mathcal{N}(0, 1)$ 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.

6 Experiments

We evaluate the proposed models using video data from the TRECVID 2003 development set [12]. 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 that shot. By removing non-informative words such as stop words and less frequent 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 5×5 regions, and extract a 15-dimensional color histogram on HVC

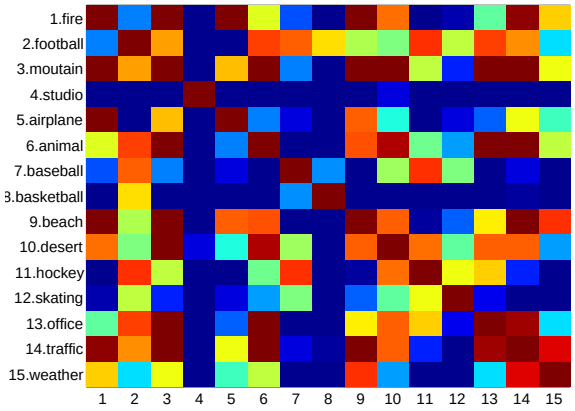


Figure 7: The color-coded matrix showing the pairwise similarity between categories revealed in the HH model. Best viewed with color.

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. First, we show some illustrative examples of the latent semantic topics derived by the proposed models and discuss the insights they provide about the structure and relationships of video categories. In the second part, we evaluate the performance of our models in video classification in comparison with some of the existing approaches.

6.1 Interpretation of latent semantic topics All the three proposed models provide intermediate representation in the form of latent semantic topics automatically derived from video data. Due to the difference on model structure, the latent topics derived from these models carry different interpretations of the same data. For both DWH and HH, the latent topics are derived from *all* the data despite their categories; therefore, they are supposed to represent some "common topics" in this data collection. In Figure 6, we show the video keyframes and keywords of the video shots that most tightly associated with 5 latent topics (i.e., having the highest conditional probability $p(h_k|\mathbf{x}, \mathbf{z})$) derived by the DWH or HH model. We find that these topics roughly correspond to some of the 15 manually defined categories. For example, in Figure 6(a), the topics are about "weather", "basketball", "airplane", "anchor", and in Figure 6(a) the topics are "studio", "baseball or football", "weather", "airplane or skating", "animal". This shows that the derived latent semantic topics are able to capture the semantics of video data.

We should also note that, since these latent topics are derived by jointly modeling the textual and imagery features of video, 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 in Figure 6(a) are very similar to each other visually; the shots of Topic 3 are not so similar visually, but it is clear that they have very close semantic meanings and share common keywords such as "flying" and "engine". A close examination also shows that the latent topics of HH have slightly better correspondence with the categories, while the latent topics of DWH seem to be clusters based on image and keyword features. This echoes with the fact that the latent topics from HH are "supervised" by the category information while those from DWH are not.

Another advantage of HH, as we discussed in Section 4.5, is that it reveals of the relationships between different categories through the hidden topics. We can tell 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

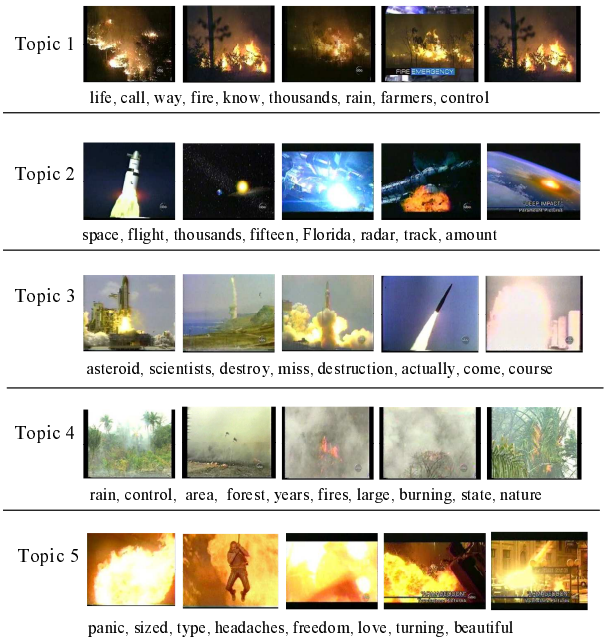


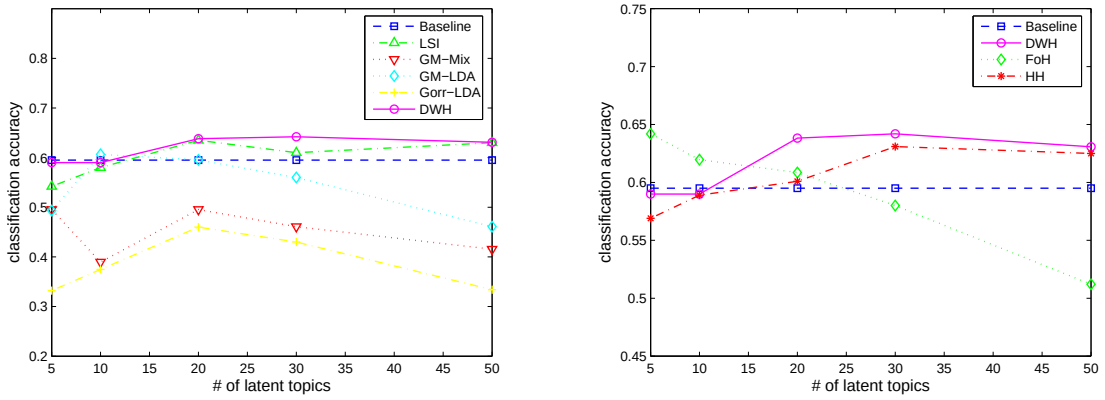
Figure 8: An illustrated of 5 latent topics of the component harmonium for "Fire" category in the FoH model.

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 7, 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.

The latent topics in a FoH model have very different interpretations. Since each component harmonium in FoH is learned independently from the data of a specific category, the latent topics of that harmonium capture the structure or sub-categories of that particular category. In Figure 8, we show the representative key-frames and keywords associated with the 5 latent topics learned from the category "Fire". We see that these 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".

6.2 Performance on video classification To evaluate the performance of DWH, 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 5, in order to examine their efficiency and accuracy. The FoH and HH model can be directly used for classification, while for the DWH model we train separate SVM classifiers based on the intermediate representation derived from DWH. 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. We implemented a baseline method which builds a SVM classifier based on the same keyword and color features used as input of our models. This is essentially an "early fusion" method since it concatenates both features into a longer feature vector.



(a) (b)

Figure 9: Comparison of classification performance (1) between various representation models (directed and undirected), and (2) between the three proposed harmonium models.

We also implemented several directed topic models, which produce intermediate representation of some kind for video data. These models include Gaussian multinomial mixture model (GM-Mixture), Gaussian multinomial latent Dirichlet allocation (GM-LDA), and correspondence latent Dirichlet allocation (Corr-LDA). The details of them models can be found in [1]. Note that all these directed models are used only for data representation, and separate SVM classifiers are trained for to perform classification based on the intermediate representations derived from these models. To make these methods comparable, we guarantee that the same kernel function and parameters are used in the SVM classifiers trained on top of different models. We use RBF kernel [3] and the best empirical parameters found by cross-validation. Also, to make the experiments tractable on various models with different learning algorithms and different numbers of latent topics, we restrict this part of experiments to a subset of our collection with the 5 largest categories containing totally 1078 shots as *airplane*, *basketball*, *baseball*, *hockey*, and *weather*.

Figure 9(a) shows the classification accuracies of all the representation models, including the undirected DWH model and the directed ones such as 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 uses the expectation-maximization (EM) method. All the models are evaluated with the number of latent semantic topics set to 5, 10, 20, 30, and 50, in order to study the relationship between performance and model complexity. Figure 9(b) compares the classification performance of the three our models, among which DWH is representation-only model while FoH and HH are classification models.

Several interesting observations can be drawn from Figure 9. First, the three undirected models as FoH, HH, and DWH achieve significantly higher performance than the directed models as GM-Mixture, GM-LDA, and Corr-LDA, which indicates 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 comparable with 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

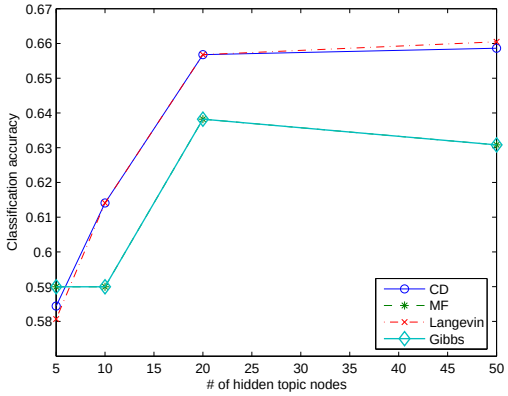


Figure 10: Classification performance of different approximate inference methods in HH

as the number of latent topics increases, which agrees with our intuition 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 it is a good property of FoH to achieve 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 10 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 perform similarly, but are slightly better than mean-filed (MF) and Gibbs sampling. We also study the efficiency of these inference methods by examining the time they need to reach convergence during training. The results show 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.

7 Conclusion

We have described three undirected graphical models for semantic representation and classification of video data. The proposed 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. Several approximate inference algorithms have been examined in terms of efficiency and classification performance.

Our HH model 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] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[4] 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.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction.*

[6] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[7] 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.

[8] M. I. Jordan. *Learning in Graphical Models: Foundations of Neural Computation.* The MIT press, 1998.

[9] 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*, pages 524–531, 2005.

[10] 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.

[11] 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.

[12] 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.

[13] P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. pages 194–281, 1986.

[14] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis, 2005.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] E. Xing, R. Yan, and A. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *Proc. of the 21th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-05)*. AUAI press, 2005.

[20] 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 dual-wing-harmonium model. We start by introducing the *general form* of exponential-family harmonium [16] 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:

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

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 $\mathbf{h}$:

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

and similarly we can derive:

$$p(\mathbf{x},\mathbf{h}) \propto \exp\Big\{ \sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{kc} \lambda_{kc} g_{kc}(h_k) + \sum_j B_j(\{\hat{\eta}_{jb}\}) \Big\}$$

$$p(\mathbf{z},\mathbf{h}) \propto \exp\Big\{ \sum_{jb} \eta_{jb} g_{jb}(z_j) + \sum_{kc} \lambda_{kc} e_{kc}(h_k) + \sum_i A_i(\{\hat{\theta}_{ia}\}) \Big\}$$

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:

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

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

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

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

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

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.(4.2), (4.3), and (4.4) are all exponential distributions. They can be mapped to the general forms above if we make the following definitions:

$$\begin{aligned}
&f_{i1}(x_i) = x_i \\
&\theta_{i1} = \alpha_i, \hat{\theta}_{i1} = \alpha_i + \sum_k W_{ik} h_k \\
&g_{j1}(z_j) = z_j, g_{j2}(z_j) = z_j^2 \\
&\eta_{j1} = \beta_j, \eta_{j2} = -1/2, \hat{\eta}_{j1} = \beta_j + \sum_k U_{jk} h_k \\
&e_{k1} = h_k, e_{k2} = h_k^2 \\
&\lambda_{k1} = 0, \lambda_{k2} = -1/2, \hat{\lambda}_{k1} = \sum_i W_{ik} h_k + \sum_j U_{jk} h_k
\end{aligned}$$

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:

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

which is exactly the same as Eq.(4.5) except the latter one is defined for a specific category.