News Video Semantic Topic Mining Based on Multi-wing Harmoniums Model

Xin Wen Xu and Yu Bo Shen
Department of traffic and transportation engineering
National University of Defense Technology
Changsha, China
{xinwen_xu, shenyubo}@126.com

Guo Hui Li
Department of System Engineering
National University of Defense Technology
Changsha, China
guohli@nudt.edu.cn

***Abstract*—Two-layer undirected graphical model, Harmoniums, is a new approach to mine latent semantic topics from observed data. For the multi-modal heterogeneous features of news video, this paper proposes multi-wing Harmoniums (MWH) model that represents news video stories as latent semantic topics derived by jointly modeling the transcript text, color histogram and edge histogram of the video. This model includes a multivariate Poisson distribution and two multivariate Gaussian distributions. It extends and improves earlier models based on two-layer random fields, which capture bidirectional dependencies between hidden topic aspects and observed inputs. The model especially facilitates efficient inference and robust topic mixing, and provides high flexibilities in modeling the latent topic spaces. The variational algorithm efficiently reduces the difficulty of model learning. The experiments results on the CCTV news video collections and an extensive comparison with various extant models show the efficiency of MWH on news video semantic mining.**

Keywords-news video multi-wing Harmoniums; video mining; semantic mining; news video

I. INTRODUCTION

Along with the rapid development of processor speed and the Internet as well as the availability of inexpensive massive digital storages, there have been strong demands for the modeling and mining of the multiple media sources, like text, image, audio and video. Numerous researchers have shown great interests to the news video as a mass medium for its easiness to acquire and richness in information. To fill the semantic gap between the low-level features and the high-level semantic topic of news video, it's necessary to make full use of the rich information provided by the multi-modal heterogeneous data so that we can effectively achieve the data mining tasks on news video like classification, cluster, retrieval and image annotation. The multi-modal heterogeneous data refers to the data of the objects to be described collected from different approaches or perspectives. And we refer to each of the approaches or perspectives as a modality. For example, in the multi-modal face detection, the multi-modal data can consist of the 2d face images and 3d face shape model; in the mining of multi-modal videos, videos can be decomposed into subtitles, audios, images, etc. Therefore, the key issues of video semantic mining researches are to model the associated data from multiple sources jointly and explore appropriate lower dimensional latent representations of the originally high-dimensional features. The fusion of multi-modal data like key frame image, audio and transcript text, has been a widely used technique in video processing. The fusion strategy includes the feature-level fusion in earlier stage and the later decision-level fusion. It is an open question as to which fusion strategy is more proper for a task. Snoek et al. [1] compares the two strategies in the classification of videos.

There are many approaches to obtain low-dimensional intermediate representations of video data. Principal component analysis (PCA) [2] 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) [3] and Fisher Linear Discriminant (FLD) [4] are also widely used in dimensionality reduction. Recently, there are also many proposals on modeling the latent semantic topics of the text and multimedia data. For example, Latent Semantic Indexing (LSI) [5] finds a linear transform of word counts into a latent eigenspace of document semantics. Though LSI can roughly obtain the latent semantic and work well in automatic index application, it generates overfitting for failing to meet the statistics principles. Later, LSI is extended to probabilistic Latent Semantic Indexing (pLSI) [6], which models the latent topic into the probability distribution of words and the documents into the probability distribution of the latent topic. The pLSI is based on the principle of probability and defines proper generative model, so it can be applied to model composition and can control the complexity. The Latent Dirichlet Allocation (LDA) [7] 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 is later extended to Gaussian-Mixture LDA and Correspondence LDA [8], both of which are used to model annotated data such as captioned images or video with transcript text.

In fact, the methods mentioned above are mainly used to transform high-dimension of raw features to low-dimensional presentation and presumably gain the latent semantics of data. However, these methods are mainly applied to single modal data and can't or can hardly be applied to the multi-modal heterogeneous data. Two-layer undirected graphical model, Harmoniums [9], is a new approach to mine latent semantic topics from observed data [16]. Based on Harmoniums models, we present news video

ISBN: 978-1-61208-265-3

multi-wing harmoniums (NVMWH) model that represents story unit as latent semantic topics derived by jointly modeling the transcript keywords, color histogram and edge histogram features of news video data for news video semantic topic mining.

The rest of the paper is structured as follows. In Section 2, we present the latent semantic topic model of news video-based on multi-wing harmoniums model and the Learning and Inference of its parameters. Section 3 presents the experiments and discussions. The paper concludes in Section 4.

II. THE LATENT SEMANTIC TOPIC MODEL OF NEWS VIDEO BASED ON MULTI-WING HARMONIUMS MODEL

A. The basic Harmoniums model

The basic Harmoniums model, which was originally studied by Smolensky (1986) [9] in his Harmony theory, defines a complete bipartite undirected graphical model containing two layers of nodes (Fig. 1). Let $H=\{h_j\}$ denotes the set of hidden units in such a graph, and let $X=\{x_i\}$ denotes the set of input units. The Harmoniums model creates a random field for the undirected graph model.

$$p(x,h|\theta)=\frac{1}{Z(\theta)}\exp\{\sum\theta_{\mathrm{i}}\phi_i(x_i)+\sum\theta_{\mathrm{j}}\phi_j(h_j)+\sum\theta_{\mathrm{ij}}\phi_{ij}(x_i,h_j)\} \qquad (1)$$

where $\phi_e(\cdot)$ denotes the potential function defined on either a singleton or a connected pair of units (indexed by e) in the model, θ_e denotes the weight of the corresponding potential, and $Z(\theta)$ stands for the log-partition function.

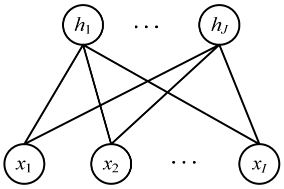

Figure 1. The basic Harmoniums Model

The bipartite topology of the harmoniums graph suggests that if the nodes within the same layer are given then the nodes of the other opposite layer are conditionally independent. This makes possible a feasible definition of the Harmoniums distribution based on the conditional distribution function $p(x|h)$ and $p(h|x)$ between the two layers: $p(x|h)=\prod_i p(x_i\,|\,h)$, $p(x|h)=\prod_j p(h_j\,|\,x)$. Hence, it is semantically simple and easy to design. For simplicity, all conditional probabilities considered here adopt exponential forms.

$$p(x_i\,|\,h)=\exp\{\sum_a\hat{\theta}_{ia}f_{ia}(x_i)\text{-}A_i(\{\hat{\theta}_{ia}\})\} \qquad (2)$$

$$p(h_i\,|\,x)=\exp\{\sum_a\hat{\lambda}_{jb}g_{jb}(h_j)\text{-}B_j(\{\hat{\lambda}_{jb}\})\} \qquad (3)$$

where $\{f_{ia}(\cdot)\}$ and $\{g_{jb}(\cdot)\}$ respectively denote the sufficient statistics of variable x_i and h_j . $A_i(\cdot)$ and $B_j(\cdot)$ denote respective log-partition functions. And the shifted parameters $\hat{\theta}_{ia}$ and $\hat{\lambda}_{jb}$ are defined as $\hat{\theta}_{ia}=\theta_{ia}+\sum_{jb}W_{ia}^{jb}g_{jb}(h_j)$ and $\hat{\lambda}_{jb}=\lambda_{jb}+\sum_{ia}W_{ia}^{jb}f_{ia}(x_i)$. It's produced by the input and all the matching pairs in hidden layers. Welling et al. [10] showed that these easily comprehensible and manipulable local conditionals precisely map to the Harmoniums random fields:

$$\begin{aligned}p(x\,|\,h)\propto\exp\{&\textstyle\sum_{\mathrm{ia}}\theta_{\mathrm{ia}}f_{ia}(x_i)+\sum_{jb}\lambda_{jb}g_{jb}(h_j)\\&+\textstyle\sum_{ijab}W_{ia}^{jb}f_{ia}(x_i)g_{jb}(h_j)\}\end{aligned} \qquad (4)$$

θ_{ia}, λ_{jb} and W_{ia}^{jb} are the set of parameters associated with their corresponding potential functions. It's very difficult to make parameter estimation for the appearance of the log-partition function with joint probability, so ratio symbol rather than precise equal symbol is used in the formula. Such a model was referred to as the Exponential Family Harmoniums (EFH) [10]. In the sequel, we will take advantage of this bottom-up strategy to construct specific Harmoniums from the easily comprehensible local conditionals. It can be shown that there is no marginal independence for either input or hidden variables in a Harmoniums model. However, an EFH enjoys the advantages of conditional independence between hidden variables, which is generally violated in the directed models. This property greatly reduces reasoning difficulty. But typically, learning harmonium is more difficult due to the presence of a global partition function.

B. The multi-wing Harmoniums model

The hidden and input units in a Harmoniums model are symmetrical, which cannot contribute to explain their causal relationship in semanteme. However, the definition mentioned above of the local condition independence based on two layers can provide explanations for the bidirectional causality of Harmoniums structure. Essentially, the hidden unit H can be considered as latent topic which defines the production of input. Conversely, H can also be seen as the predictors produced by a discriminative model of the input unit.

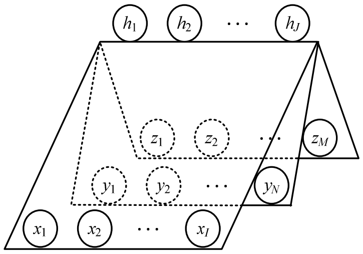

Figure 2. Multi-wing Harmoniums Model

In many applications, the input to the model does not have to be from a single source and/or of a homogeneous data type. For example, in the analysis of typical multi-media application video streaming, the input from video clips contains much relevant information, such as transcript texts, pictures, sounds and motion vectors. Assuming that all these inputs are combined together to present one topic center, it will be natural to model the shared topic center using a set of hidden units, and to group observations from all sources into multiple homogeneous arrays of input units, each corresponding to a single source. Thus a multi-wing Harmoniums model is constructed, as shown in Fig. 2. This model consists of three canonical Harmoniums joint by a shared array of hidden units. It's straightforward to construct a multi-wing Harmoniums model from a canonical Harmoniums model. For example, a three-wing Harmoniums model added by two input sets *Y*={y_n} and *Z*={z_m} can be related to *H* via $p(y\,|\,h)=\Pi_n\, p(y_n\,|\,\mathrm{h})$ and $p(z\,|\,h)=\Pi_m\, p(z_m\,|\,h)$, where

$$p(y_n\,|\,h)=\exp\left\{\sum_{\mathrm{c}}\hat{\gamma}_{nc}e_{nc}(y_n)-C_n(\hat{\gamma}_{nc})\right\} \tag{5}$$

$$p(z_m\,|\,h)=\exp\left\{\sum_{\mathrm{d}}\hat{\eta}_{md}s_{kd}(z_m)-D_m(\hat{\eta}_{md})\right\} \tag{6}$$

$$\hat{\gamma}_{nc}=\gamma_{nc}+\sum_{jb}U_{nc}^{jb}g_{jb}(h_j) \tag{7}$$

$$\hat{\eta}_{md}=\eta_{md}+\sum_{jb}V_{md}^{jb}g_{jb}(h_j) \tag{8}$$

Together with (2), and slightly modified (3) that takes into account the influences from *X*, *Y*, and *Z* by loading the parameter $\hat{\lambda}$ with additional shift

$$\hat{\lambda}_{jb}=\lambda_{jb}+\sum_{ia}W_{ia}^{jb}f_{ia}\left(x_i\right)+\sum_{nc}U_{nc}^{jb}e_{nc}\left(y_n\right)+\sum_{md}V_{md}^{jb}s_{md}\left(z_m\right) ,$$

where $\{U_{nc}^{jb}\}$ and $\{V_{md}^{jb}\}$ stand for the matching parameters between the hidden unit and the input sets *Y* and *Z*. Thus, we can get the random field of three-wing exponential family Harmoniums.

$$\begin{aligned} p\left(x,y,z,h\right)\propto& \\ &\exp\{\sum_{ia}\theta_{ia}f_{ia}\left(x_i\right)+\sum_{nc}\gamma_{nc}e_{nc}\left(y_n\right)+\sum_{md}\eta_{md}s_{md}\left(z_m\right) \\ &+\sum_{jb}\lambda_{jb}g_{jb}\left(h_k\right)+\sum_{ijab}W_{ia}^{jb}f_{ia}\left(x_i\right)g_{jb}\left(h_k\right) \\ &+\sum_{njcb}U_{nc}^{jb}e_{nc}\left(y_n\right)g_{jb}\left(h_k\right)+\sum_{mjdb}V_{md}^{jb}s_{md}\left(z_m\right)g_{jb}\left(h_k\right)\} \end{aligned} \tag{9}$$

where $\{f_{ia}(\cdot)\}$, $\{e_{nc}(\cdot)\}$, $\{s_{md}(\cdot)\}$, $\{g_{jb}(\cdot)\}$ stands for the fully statistical characteristics of x_i, y_n, z_m, h_k. This construction maintains the conditional independence between hidden variables given inputs and hence ensures the efficiency of inference (once the model is parameterized). Note that *x*, *y*, *z* are marginally dependent to each other, as can be quickly verified from the bipartite graph structure. This enables the application of inferring the features of one source from anther source, for example, automatic image annotation which attempts to infer related words from a given image.

C. The multi-wing Harmoniums model of news video

To model video streams, which contain both text and image information, in the following, we outline a news video multi-wing harmonium (MWH) model based on a text submodel and two image submodels using the modular constructive technique described above.

The parameters of this model are defined as follows:

- The story unit of news video *s* denotes a four-dimensional tuple of (x,y,z,h) , which respectively denotes the keywords, color features and texture features of key-frame, and latent semantic topics.
- The vector $x=(x_1,x_2,\cdots,x_I)$ denotes the keyword feature extracted from the transcript associated with the shot. Here *I* is the size of the key words dictionary, and $x_i\in\{0,1\}$ indicates the absence or presence of the keyword *i* in the story of news video.
- The vector $y=(y_1,y_2,\cdots,y_N)$ denotes the boundary histogram feature of the key frame in the story of news video. Similar to the color feature, each key frame is divided into rectangular regions of size *N* in fixed size. $y_n\in R^C$ presents the color histogram feature of the region of size N denoted by C-dimensional vector. So y is a stack vector whose length is *CN*.
- The vector $z=(z_1,z_2,\cdots,z_N)$ denotes the color histogram feature of the key frame in the story of news video. Each key frame is evenly divided into rectangular regions of size *N* in fixed size. $z_n\in R^C$ presents the color histogram feature of the region of size *N* denoted by D-dimensional vector. So z is also a stack vector whose length is *DN*.
- The vector $h=(h_1,h_2,\cdots,h_J)$ denotes the latent semantic topics of the story of the news video. *J* is the number of latent topic, $h_j\in R$ denotes the degree of correlation between the news story and the latent topic of the size *j*.

1) Text feature model

Following the traditional bag-of-word model [11] for texts, we model the video transcript texts by adopting the word-count Poisson distribution model of the document. Instead of using a continuous surrogate of the discrete counts (as done in a mixture of Gaussian setting), or assuming that the counts of words are accumulated from independent draws from multinomial distributions (as done in LDA), the text Poisson model is based on the hypothesis that the rate of the word in a document can be described as the word's accumulation of the Poisson distribution in the dictionary, namely the latent topic features associated with each document directly determine the expected rate of each word in a document. In this way, a Poisson distribution is assigned to the observation counts of each word in a document. This text model has key differences from a multinomial model, which is achieved directly by topic mixing in the distribution of word rates in the documents combining specific topic feature, rather than via an additive effect of multiple single topic draws of the same word or via marginalization of the latent topic of each word. In the conditional Poisson model for word counts, topic mixing is still stable and robust even when a word appears only once or a few times in a document, which is typical in video captions though. Whereas in the multinomial word model, single word *i* can only come from a single topic and is thus unable to satisfy the topic mixing directly. The text Poisson submodel is defined as follows: For each word $i=\{1,\cdots,I\}$, its rate x_i is distributed as:

$$p(x_i \mid h) = \mathrm{Poisson}(x_i | \exp(\alpha_i + \sum_j h_j W_{ij})) \qquad (10)$$

This model shows that, each key word of the transcript text in the story of news video relies on a Poisson distribution of the latent semantic topic *h*. In other words, the probability of a key word's appearance is determined by the weighted array of latent topic feature *h*. The parameters α_i and W_{ij} are scalar variables. $\alpha = (\alpha_1, \cdots, a_I)$ is a I-dimension vector. $W = [W_{ij}]$ is a matrix of $I \times J$. Because of the conditional independence between x_i and *h*, there is $p(x \mid h) = \prod_i p(x_i \mid h)$.

2) Image feature model

As to the image feature of the story in news video, we denote the feature by adopting the 72-dimension feature of color histogram and the 80-dimension of edge histogram in the partition HSV of key frame image in video stories. The color histogram feature y_n of the zone *N* in key frame image and the edge histogram feature z_n separately satisfies the distribution of conditional multivariate Gaussian distribution.

$$p(y_n \mid h) = N(y_n \mid \sigma_n^2(\beta_n + \sum_j h_j U_{nj}), \sigma_n^2) \qquad (11)$$

$$p(z_n \mid h) = N(z_n \mid \mu_n^2(\tau_n + \sum_j h_j V_{nj}), \mu_n^2) \qquad (12)$$

where β_n and U_{nj} are 72-dimension vectors. $\beta = (\beta_1, \cdots, \beta_N)$ is a stack vector of 72×*N*. $U = [U_{nj}]$ is a matrix of 72×*N*×*J*, and σ_n^2 is a 72×72−dimension covariance matrix. τ_n and V_{nj} are 80-dimension vectors， $\tau = (\tau_1, \cdots, \tau_N)$ is a stack vector of 80×*N*. $V = [V_{nj}]$ is a matrix of 80×*N*×*J*, and μ_n^2 is a 80×80−dimension covariance matrix. We adopt unit matrix to simplify operation. For the conditional independence between yn, zn and h, there are $p(y|h) = \prod_n p(y_n \mid h)$ and $p(z \mid h) = \prod_n p(z_n \mid h)$.

3) Hidden semantic topic model

As to the hidden unit *h* of latent topic feature in the news video story unit, we assume that each feature is a conditional unit-variance Gaussian distribution, whose mean is determined by a weighted combination of the key word feature in the video news story unit's transcript text, the color histogram and the edge histogram.

$$p\left(h_j \mid x, y, z\right) = N\left(h_j \mid \sum_i W_{ij} x_i + \sum_n U_{nj} y_n + \sum_m V_{mj} z_m, 1\right) \qquad (13)$$

where W_{ij}, U_{nj} and V_{mj} share the same parameters with (10), (11) and (12). Similarly there is $p(x \mid h) = \prod_i p(x_i \mid h)$ from the conditional independence. This model denotes the topic vector as a random point in Euclidean space, while other feature models based on polynomial denote their topic joint vector as a point in the space of single feature. The condition distribution of all vectors in the model is shown above. These local conditional distributions can map into the random filed of Harmoniums shown as follows.

$$p\left(x, y, z, h\right) \propto \exp\{\sum_i \alpha_i x_i + \sum_n \beta_n y_n + \sum_n \tau_n z_n - \sum_n \frac{y_n^2}{2} - \sum_n \frac{z_n^2}{2} - \sum_j \frac{h_j^2}{2} + \sum_{ij} W_{ij} x_i h_j + \sum_{nj} U_{nj} y_n h_j + \sum_{nj} V_{nj} z_n h_j\} \qquad (14)$$

By integrating out the hidden variables h in (14), we obtain the marginal distribution over the observed keyword and color features in the stories of news video.

$$p\left(x, y, z\right) \propto \exp\{\sum_i \alpha_i x_i + \sum_n \beta_n y_n + \sum_n \tau_n z_n - \sum_n \frac{y_n^2}{2} - \sum_n \frac{z_n^2}{2} + \frac{1}{2}\sum_j (\sum_i W_i x_i + \sum_n U_{nj} y_n + \sum_n V_{nj} z_n)^2\} \qquad (15)$$

which also contains a hidden partition function in this distribution.

The parameter of the NVMWH model, $s = (\alpha, \beta, \tau, W, U, V)$, is learnt by the maximum likelihood of a news video story set, where the likelihood function is defined by (15). Due to the presence of the global partition function, the learning process requires approximate inference methods, which will be discussed in the next section. 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 $p\left(h_j \mid x_i, y_n, z_n\right)$, but it would not lead to more general representations in terms of probability $p(x, y, z)$ [10].

From the analyses above, we can find that the NVMWH model can denote the latent semantic topic of the story units in news video. In other words, the NVMWH model can be used to infer the latent semantic topic h if given the text feature x, the visual feature y and z of key frame image of the story unit in the news video.

D. The learning of the model's parameter

By the analysis in the section above, we can find that the NVMWH model can be used to gain the hidden semantic topic of the story unit in news video but the parameters of the model have to be determined before using this model. There are many methods to estimate the model parameters. We adopt the maximum likelihood method according to the NVMWH model defined in the section above. Assuming that the training set contain *N* independent identically distributed (IID) story units, namely $\{x, y, z\} = \{x_n, y_n, z_n, n = 1, \cdots, N\}$, the parameter of the NVMWH model $s = (\alpha, \beta, \tau, W, U, V)$ is estimated by maximizing the log-likelihood of the data defined by (15). Due to the complexity of this 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 (15) with respect to model parameters:

$$\delta\alpha_i = \left\langle x_i \right\rangle_{\tilde{p}} - \left\langle x_i \right\rangle_p, \delta\beta_n = \left\langle y_n \right\rangle_{\tilde{p}} - \left\langle y_n \right\rangle_p,$$
$$\delta z_n = \left\langle z_n \right\rangle_{\tilde{p}} - \left\langle z_n \right\rangle_p, \delta W_{ij} = \left\langle x_i h'_j \right\rangle_{\tilde{p}} - \left\langle x_i h'_j \right\rangle_p, \qquad (16)$$
$$\delta U_{nj} = \left\langle y_n h'_j \right\rangle_{\tilde{p}} - \left\langle y_n h'_j \right\rangle_p, \delta V_{nj} = \left\langle z_n h'_j \right\rangle_{\tilde{p}} - \left\langle z_n h'_j \right\rangle_p$$

where $h'_j = \sum_i W_i x_i + \sum_n U_{nj} y_n + \sum_n V_{nj} z_n$, $\langle \cdot \rangle_p$ and $\langle \cdot \rangle_{\tilde{p}}$ denote expectation under empirical distribution (i.e., data average) or model distribution of the harmonium, respectively. Like other undirected graph models, there is global normalizer (a.k.a partition function) in the likelihood function of the

NVMWH model, so it's very difficult to directly calculate $\langle\cdot\rangle_p$. Instead, we need approximate inference methods to estimate these model expectations $\langle\cdot\rangle_p$. We explored three approximate inference methods in our work, which are briefly discussed below.

1) The 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 [12]. We use the naive version of MF, where the joint probability *p* is approximated by a surrogate distribution *q* as a product of singleton marginals over the variables:

$$q(x,y,z,h) = \prod_i q(x_i \left| \iota_i \right.) \prod_n q(y_n \left| \varsigma_n, \sigma_n \right.) \\ \prod_n q(z_n \left| \zeta_n, \mu_n \right.) \prod_j q(h_j \left| \xi_j \right.) \tag{17}$$

where the singleton marginals are defined as $q(x_i) \sim \text{Poisson}(\iota_i)$, $q(y_n) \sim N(\varsigma_n, \sigma_n)$, $q(z_n) \sim N(\zeta_n, \mu_n)$ and $q(h_j) \sim N(\xi_j, 1)$, and $\{\iota_i, \varsigma_n, \zeta_n, \xi_j\}$ is variational parameter. The variation parameters can be computed by minimizing the KL-divergence between *p* and *q*, which results in the following fixed-point updating equations:

$$\iota_i = \exp(\alpha_i + \sum_j W_{ij} \xi_j) \tag{18}$$

$$\varsigma_n = \sigma_n^2 (\beta_n + \sum_j U_{nj} \xi_j) \tag{19}$$

$$\zeta_n = \mu_n^2 (\tau_n + \sum_j V_{nj} \xi_j) \tag{20}$$

$$\xi_j = \sum_i W_{ij} \iota_i + \sum_n U_{nj} \varsigma_n + \sum_n V_{nj} \zeta_n \tag{21}$$

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 (16), 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 (16). Besides, when using the gradient ascent method in the mean field, the learning process includes two nested loops: the outer loop iteratively updates the model parameters using the learning rules (16), 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*.

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 [13] [14]. 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. If the iteration number is big enough, the sampling of joint distribution and the boundary distribution is gained successively. For example, in the Poisson text submodel mentioned in this paper, the sampling order is defined as $x_1, \cdots, x_I, h_1, \cdots, h_J$ and other variables are defined as inputs. First, $\{h_j\}$ is set as the current value and each x_i is sampled from the condition distribution defined in (10). Then, x_i is set as the current value and each h_j is sampled from the condition distribution defined in (13) , 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.

3) Contrastive divergence

An alterative to exact gradient ascent search based on the learning rules in (16) is the contrastive divergence (CD) algorithm [15] that approximates the gradient learning rules. In each step of the gradient update, 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 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. 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.

III. Experiments and Discussions

To verify the effectiveness of the NVMWH model, our experiments mainly include 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.

A. *Experimental data and features selection*

The experimental data adopted in our experiment come from the news video stories from CCTV news. We collect news programs of two years and four months, from May of 2006 to July of 2008, which contain 25776 news story units. Each story unit of the news video is considered as a document or a training test example. Our experiment adopts 4214 news story units which belong to 18 categories, namely fire disaster, flood, earthquake, storm (typhoon and hurricane), Olympic games, bird flu, Taiwan, the Korean nuclear issue, the United Nations, the United States, Russian, Japan, Iran, Iraq, terrorist attack, country, oil price and football. Each story unit is related to a category. Because the CCTV news shows only includes various important news, the distribution is uneven of the story unit in each category. The number of every kind of story unit in this experiment ranges from 26 to 828. 30% of the samples of each category are randomly selected as training samples and the rest are considered as the test samples. Table 1 describes the training and the test samples of the experiment in detail.

TABLE I. THE SAMPLE DISTRIBUTION OF TRAINING SET AND TEST SET

Serial number	Category name	Total number of samples	Number of training samples	Number of test samples
1	fire disaster	80	24	56
2	flood	61	18	43
3	earthquake	627	188	439
4	storm	106	32	74
5	Olympic games	828	248	580
6	bird flu	26	8	18
7	Taiwan	113	34	79
8	the Korean nuclear issue	38	11	27
9	the United Nations	180	54	126
10	the United States	363	109	254
11	Russian	259	78	181
12	Japan	276	83	193
13	Iran	244	73	171
14	Iraq	142	43	99
15	terrorist attack	85	26	60
16	country	643	193	450
17	oil price	114	34	80
18	football	29	9	20
Total		**4214**	**1264**	**2950**

As to the text feature, we download all the transcript texts of all the news story units from the CCTV website. We adopt the word segmentation software of Beijing Language and Culture University to conduct word segmentation, and leave out all stopwords, and merger synonyms and near-synonyms. About nine thousand key words are gained from the 18 topics above. To further reduce the complexity of the model operation, we ignore the low-frequency words whose frequencies are less than 6 and extract 3182 keywords as the text feature. Hence, the text feature of each story unit in the news video is denoted as a 3182-dimension binary feature vector where 1 stands for the appearance of a certain keyword in the story unit and 0 stands for no appearance. As to the visual feature, we adopt the 72-dimension HSV color histogram feature and the 80-dimention edge histogram feature of the key frame image in MPEG-7.

By default, NVMWH is trained via contrastive divergence with up to 1000 steps of gradient ascent. We adopt the mean field approximation and the Gibbs sampling to conduct the model training.

To mitigate the issue of “identifiability” [10] that allows multiple parameters to share the same marginal likelihood, the initial estimations of parameters *W*, *U* and *V* in the NVMWH were determined by a SVD on the design matrix of text/images features over shots. We do not strongly emphasize this issue because in our analysis NVMWH is not mainly used to directly capture the exact semantics of the latent factors underlying the data space. In order to achieve semantically more accurate and informative latent factor representations, we can apply a subsequent clustering procedure on the lower-dimensional representations provided by NVMWH.

The parameters of GM-Mix and GM-LDA were obtained using EM. We infer the latent topic captured by GM-Mix using the conditional probabilities of hidden variables $p(\mathrm{h}|\mathrm{x},\mathrm{z})$ and those by GM-LDA based on variational Dirichlet posteriors of the topic weights.

B. The results and analysis of the latent topic mining experiment

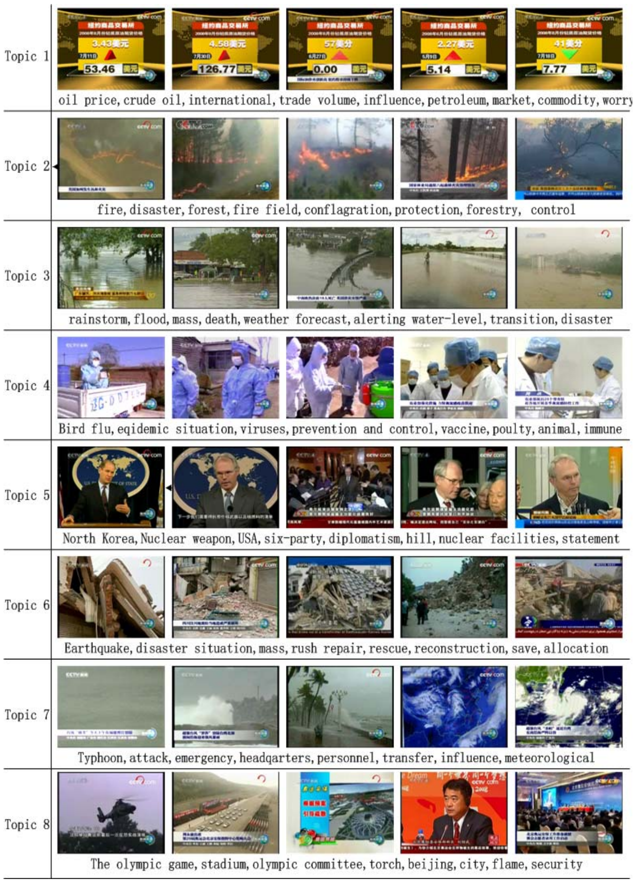
Topic 1
oil price, crude oil, international, trade volume, influence, petroleum, market, commodity, worry
Topic 2
fire, disaster, forest, fire field, conflagration, protection, forestry, control
Topic 3
rainstorm, flood, mass, death, weather forecast, alerting water-level, transition, disaster
Topic 4
Bird flu, eqidemic situation, viruses, prevention and control, vaccine, poulty, animal, immune
Topic 5
North Korea, Nuclear weapon, USA, six-party, diplomatism, hill, nuclear facilities, statement
Topic 6
Earthquake, disaster situation, mass, rush repair, rescue, reconstruction, save, allocation
Topic 7
Typhoon, attack, emergency, headqarters, personnel, transfer, influence, meteorological
Topic 8
The olympic game, stadium, olympic committee, torch, beijing, city, flame, security

Figure 3. Examples of latent topics

The NVMWH model can automatically discover low-dimension meaningful latent topics from the high-dimension multi-modal features of the texts and images in the news videos. We illustrate 8 latent topics out of the 18 topics learned by NVMWH in Fig 3. Each topic is described by the first 8 keywords and the first 5 key frame images related to the video shots. These keywords and key frame images possess highest condition probability in the latent topics. It is shown that the first 6 topics correspond to the scenes of oil price, forest fire, flood, bird flu, the Korean nuclear and the earthquake. They are a cluster based on the transcript texts and images. The last two topics illustrate some interesting patterns discovered by NVMWH. At the first sight, these key frames of the shots show great differences in visual features.

ISBN: 978-1-61208-265-3

It seems to present several totally different topics for the different scenes like helicopter, stadium, meeting, shore and meteorological chart. However, by examining the transcript texts of the news story units, we find that their semantic topics share some common aspects. Several news stories in topic 7 all mention the similar or same words, such as typhoon, attack and weather which are words related to the topic typhoon. Similarly, several news stories in topic 8 also mention some related words like Olympic Games, stadium and security work. Obviously, NVMWH model discovers the last two topics mainly based on the similarity between the key words of the transcript texts in the videos, while the visual features functions a lot in discovering other topics.

C. The results and analysis of the Classification performance experiment of NVMWH model

In order to show the predictive power of the low-dimension latent semantic topic produced by NVMWH, we adopt classification, which is the most important task in multimedia analysis and application, to evaluate the performance of the NVMWH. In the experiment, the dimension of the latent semantic topic is set as less than 50 and its original feature dimension is set as 3182-dimension. The color histogram feature of the key frame image is set as 72-dimension and the edge histogram feature is set as 80-dimension.

First, we evaluate the performance of NVMWH on classifying testing examples into one of the predefined categories, and compare this method with LSI, GM-Mix and GM-LDA. For each algorithm, the parameters are estimated using all data, ignoring their true class labels. Once the models are learned, we use them to project every example into a lower-dimensional latent topic space. Then we split the data into a training set and a testing set as shown in Table 1, use the SVM^{Light} package to learn a support vector machine (SVM) on the training data, and predict on the testing data.

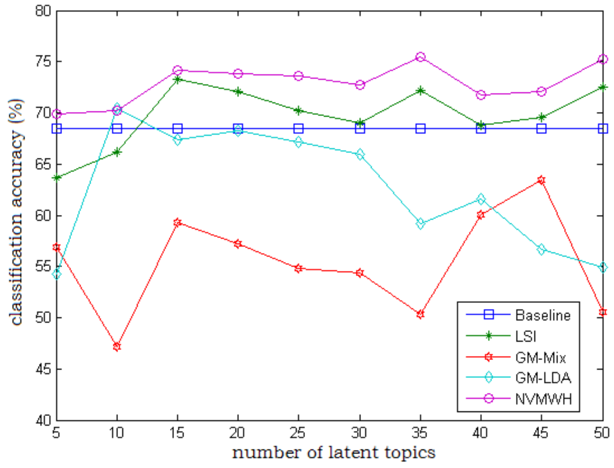

Figure 4. The classification accuracy of every model at different topic dimension

Fig. 4 shows the classification accuracy of five different models at different latent topic feature dimension ranging from 5 to 50, and the dimension is the number of the latent semantic topics. The Baseline method retains the available feature classification results of all the variables in training and test. We find that the NVMWH model can always achieve higher classification accuracy than the Baseline even with a large dimensionality reduction. Under the same topic dimension, it also outperforms LSI with a good margin. We believe that this may be partially explained by the arguably better assumptions adopted by NVMWH on modeling text/image features. Surprisingly, GM-Mix produces a considerably worse performance than the baseline because the modeling power of GM-Mix is too limited to capture multiple latent topics for each text/image pair. Too much information is eliminated in GM-Mix's representations because the posterior distribution is usually peaked at one latent topic. Compared to GM-Mix, GM-LDA offers more flexibilities in modeling associated text/images and indeed it is (slightly) superior to all other models when latent aspect dimension is set to be 10. But it appears that GM-LDA may have suffered from overfitting or a low-dimensionality bias as its error curve rises significantly in higher-dimensional latent space. In contrast, we observe that the performances of LSI and NVMWH are relatively stable over a wide range of dimensions, which may reflect the robustness and expressiveness of their representation schemes for the latent aspects (i.e., as Gaussian variables rather Dirichlet variables).

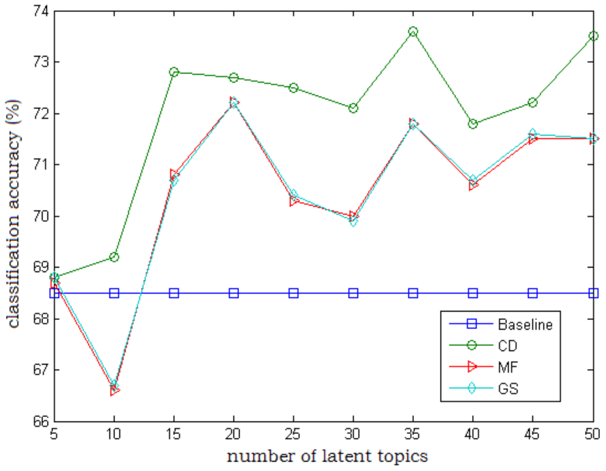

Figure 5. The classification accuracy of different leaning algorithms in the NVMWH model

Fig. 5 presents the comparison of the performance between mean field, Gibbs sampling and contrastive divergence. We discover that mean field and the Gibbs sampling are similar in performance, while the accuracy of contrastive divergence is slightly better than the two. It's mainly because the latter approach uses a fully factorized distribution to approximate the true distribution whereas the former uses a Monte Carlo approximation. Meanwhile, we make a study of the efficiency of the three methods by examining the time taken to reach the convergence of the learning methods during the training. The results show that the mean filed possess the highest efficiency at 2 minutes, the contrastive divergence method ranks second at 10 minutes and the Gibbs sampling has the lowest efficiency for about 50 minutes. Therefore, it is better to adopt the contrastive divergence method to conduct the learning and inference of the model parameters for the comprehensive consideration of accuracy and efficiency.

On the same news subject, why does not the classification performance increase steadily corresponding to the number of the latent topics? That is because the news subject is determined by the latent topics with high probability distribution, and the latent topics after certain dimension cannot help express the content of the subject of news. That is to say, the latent topics with low probability distribution have few associations, or even no associations, with the news subject. Thus, the classification performance would not increase with increasing number of latent topics.

IV. CONCLUSION AND FUTURE WORK

We make a thorough study of the latent semantic topic mining of news video by adopting the multi-wing two-layer undirected graphical model. First, we construct a news video multi-wing Harmoniums model of multi-modal heterogeneous features based on the basic Harmoniums model, the transcript texts and the key frame images in the news video. In this model, the multivariate Gaussian variables denoted the latent topics. The condition distribution of specific features models on the inputs of various kinds of data resources, namely a multiple Poisson distribution is used to model the text features and two multivariate Gaussian models respectively denote features of color histogram and of the edge histogram in the key frame image of news video story. The probability distributions are determined by all the topics so that a better topic mixing is achieved. It expands and improves the previous random field model, which is based on two layers by the bidirectional dependence relationship between the latent topics and the observed input data, whose performance is especially shown in the aspects of promoting effective reasoning, robust topic mixing and flexible latent topic modeling. The experiments of the latent semantic topic extraction and the prediction performance based on the NVMWH model prove the strong presentation ability and robustness of NVMWH model on latent semantic topic mining in news video stories.

In the future work, we will adopt more audio-visual features like facial feature, voiceprint feature, etc. in NVMWH model, and study a more effective learning algorithm for model’s parameters, so as to improve the mining precision and efficiency of video semantic topics and apply semantic features to the mining and analysis of intelligence in public news videos.

REFERENCES

[1] C. Snoek, M. Worring, and A. Smeulders. “Early versus late fusion in semantic video analysis,” Proceedings of 13th ACM International Conference on Multimedia (MM 2005) in Singapore, ACM Press, Nov. 2005, pp. 399-402.

[2] I. T.Jolliffe. “Principal component analysis,” 2nd ed., Springer Press, 2002.

[3] A. Hyvarinen, J. Karhunen, and E. Oja. “Independent component analysis,” New York, USA: wiley, 2001.

[4] L. Devroye, L. Györfi, and G. Lugosi, “A probabilistic theory of pattern recognition,” Springer Press, 1996, pp. 46-47.

[5] S. C. Deerwester, S. T. Dumais, and T. K. Landauer. “Indexing by latent semantic analysis,” Journal of the American Society of Information Science, Sep. 1990, 41(6) pp. 391-407.

[6] T. Hofmann. “Probabilistic latent semantic analysis,” Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Jul. 1999, pp. 289-296.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent dirichlet allocation,” Journal of Machine Learning Research, MIT Press, Mar. 2003, pp. 993-1022.

[8] D. M. Blei and M. I. Jordan. “Modeling annotated data,” Proceedings of the 26th annual international ACM SIGIR conference on Research and decvelopment in information retrieval, ACM, Jan. 2003, pp. 127-134.

[9] P. Smolensky. “Information processing in dynamical system: foundations of harmony theory. Parallel distributed processing: explorations in the microstructure of cognition foundations”, Cambridge: MIT Press. 1986, pp. 194-281.

[10] M. Welling, M. Rosen-Zvi, and G. Hinton. “Exponenetial family Harmoniums with an application to information retrieval,” In Advance Neural Information Processing Systems, Dec. 2005, pp. 1481-1488.

[11] D. Metzler. “Beyond bags of words: effectively modeling dependence and features in information retrieval,” SIGIR Forum, Dec. 2008, 42(1), pp. 77−77.

[12] 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, Aug. 2003, pp. 583-591.

[13] W. R. Gilks, S. Ripley, and D. J.Spiegelhalter. “Markov chain Monte Carlo in practice,” Cambridge, UK: Chapman & Hall/CRC Press, 1996.

[14] J. R. Smith and S. F. Chang. “Visually searching the web for content,” IEEE Multimedia Magazine, Jul.-Sep. 1997, 4(3), pp. 12-20.

[15] M. Welling and G. E. Hinton. “A new learning algorithm for mean field Boltzmann machines,” Proceedings of the International Conference on Artificial Neural Networks, London, UK, Springer-Verlag, Aug. 2002, pp. 351-357.

[16] E. Xing, R. Yan, and A. Hauptmann. “Mining associated text and images with dual-wing harmoniums,” Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Jul. 2005, pp. 633-641.