



Bayesian Nonparametric Topic Models for Video Data Analysis

Xue, Jianfei

(Degree)

博士 (工学)

(Date of Degree)

2018-03-25

(Date of Publication)

2019-03-01

(Resource Type)

doctoral thesis

(Report Number)

甲第7202号

(URL)

<https://hdl.handle.net/20.500.14094/D1007202>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



Doctoral Dissertation

Bayesian Nonparametric Topic Models for Video Data Analysis

(映像データ解析のためのベイジアンノンパラメトリック・トピックモデルに関する研究)

January 2018

Graduate School of System Informatics
Kobe University

XUE JIANFEI

薛建飛

Abstract

Bayesian topic modeling methods such as latent Dirichlet allocation (LDA) and hierarchical Dirichlet processes (HDP) are deemed as well-known machine learning or data mining approaches. For original Bayesian topic models, they usually focus on text data analysis, where a generative process of documents is modeled based on a hypothesis that text words in each document are generated from a mixture distribution of latent topics, where each latent topic is represented as a word distribution. Nonparametric topic model is a typical type of Bayesian topic models, such as HDP which models multiple correlated documents with multiple infinite Dirichlet processes that are connected by sharing a mixture of infinite components (topics) among all the documents.

In recent years, people are paying more and more attentions on extending such topic models to analyze video data. However, such modeling works are very challenging, since the composition of video data is more abstract and complicated than the one of simple text data. To be more specific, video data can be considered to be a type of multimodal data that generally contain a sequence of video frames, audio, and speech transcripts. To analyze such multimodal video data, in this paper, I propose a number of Bayesian nonparametric topic modeling methods, which mainly focus on tackling three major problems: (1) the way to model the time relativity between neighboring video frames, (2) the way to model the multimodal data within a unified topic model, and (3) the way to make the model perform a specific application on video data analysis. These nonparametric topic modeling methods are respectively developed with respect to two application scenarios of video data analysis: (1) video semantic analysis, (2) video emotion recognition. Finally, a series of comprehensive experiments are conducted to evaluate the performance of proposed methods.

Table of contents

List of figures	vi
List of tables	viii
1 Introduction	1
1.1 Background	1
1.2 Related Work	3
1.3 Outline	5
2 Preliminary Work	6
2.1 Background	6
2.2 Hierarchical Dirichlet Processes	7
2.2.1 Generative Process	7
2.2.2 Stick-Breaking Construction	8
2.2.3 Chinese Restaurant Franchise	10
2.3 Inference Method for HDP	12
3 Sequential Correspondence Hierarchical Dirichlet Processes for Video Semantic Analysis	15
3.1 Overview	15
3.2 Sequential HDP	16
3.2.1 Generative Process	16
3.2.2 Stick-Breaking Construction	18
3.3 Correspondence Method for Multimodal Data Modeling	20
3.4 Inference Method	22
3.4.1 Factors in Chinese Restaurant franchise	22
3.4.2 Posterior Representation Sampler	23
3.4.3 Cascaded Gibbs Sampling	27

3.5	Experiments	28
3.5.1	Experiment Description	28
3.5.2	Parameter Tuning	29
3.5.3	Evaluation of Models	31
3.6	Summary	35
4	Sequential Symmetric Correspondence Hierarchical Dirichlet Processes for Video Semantic Analysis	36
4.1	Overview	36
4.2	Preliminary Work on Sequential HDP	37
4.3	Multimodal Data Modeling Methods	38
4.3.1	Conditionally Independent Modeling Method	39
4.3.2	Correspondence Modeling Method	40
4.3.3	Symmetric Correspondence Modeling Method	41
4.4	Inference Method for Seq-Sym-cHDP	44
4.4.1	Posterior Representation Sampler	44
4.4.2	Cascaded Gibbs Sampling	46
4.5	Experiments	47
4.5.1	Experimental Setup	47
4.5.2	Parameter Tuning	48
4.5.3	Qualitative Evaluation	49
4.5.4	Quantitative Evaluation	54
4.6	Summary	56
5	Supervised Bayesian Nonparametric Topic Modeling Methods for Multi-class Video Classification	58
5.1	Overview	58
5.2	Supervised Sequential HDP	59
5.3	Supervised Sequential Symmetric Correspondence HDP	61
5.4	Inference Method and Class Prediction	63
5.4.1	Posterior Representation	63
5.4.2	Cascaded Gibbs Sampling	65
5.4.3	Softmax Regression and Class Prediction	66
5.5	Experiments	67
5.5.1	Experimental Setup	67
5.5.2	Qualitative Evaluation	69

5.5.3	Quantitative Evaluation	70
5.6	Summary	72
6	Bayesian Nonparametric Multimodal Data Modeling Framework for Video E-motion Recognition	73
6.1	Overview	73
6.2	Framework	74
6.3	Description of Features	76
6.3.1	CNN-based Deep Features	76
6.3.2	Extraction of Visual Words and Audio Words	77
6.4	Symmetric Correspondence HDP	77
6.4.1	Generative Process	78
6.4.2	Inference Method	80
6.5	Experiments	81
6.5.1	Experimental Setup	81
6.5.2	Unimodal Evaluation on CNN-based Deep Features	82
6.5.3	Multimodal Evaluation on Modeling Methods	84
6.6	Summary	84
7	Conclusion	86
	Acknowledgements	88
	References	89
	Publications	93

List of figures

2.1	Graphical model representation of HDP.	8
2.2	Graphical model representation for stick-breaking construction of Seq-HDP.	9
2.3	A depiction of a Chinese restaurant franchise.	11
3.1	Graphical model representation of Seq-HDP.	17
3.2	Graphical model representation for stick-breaking construction of Seq-HDP.	19
3.3	Graphical model representation for stick-breaking construction of Seq-cHDP.	21
3.4	Generation mechanism of stick-breaking construction for Seq-cHDP associated with the extracted counts of factors in CRF.	22
3.5	Result for parameter tuning on Seq-cHDP with accuracy.	29
3.6	Result for parameter tuning on Seq-cHDP with macro-F1.	30
3.7	Area graph for topic distribution in video No.109 (© <i>Span-TV</i>), superimposed with speech words.	31
3.8	Area graph for topic distribution in video No.221 (© <i>Blue Orchid Productions</i>), superimposed with speech words.	32
3.9	Area graph for topic distribution in video No.106 (© <i>Darien Library</i>), superimposed with speech words.	33
4.1	Simplified Graphical model representation for stick-breaking construction of Seq-CI-HDP.	39
4.2	Simplified Graphical model representation for stick-breaking construction of Seq-cHDP.	41
4.3	Simplified Graphical model representation for stick-breaking construction of Seq-Sym-cHDP.	42
4.4	Parameter tuning results for Seq-Sym-cHDP.	47
4.5	Performance of pivot assignments in video No.161 with Seq-Sym-cHDP.	48
4.6	Performance of pivot assignments in video No.511 with Seq-Sym-cHDP.	49

4.7	Performance of pivot assignments in video No.984 with Seq-Sym-cHDP. . .	50
4.8	Area graph for topic distribution in video No.161 (© <i>The Reality Report</i>) with Seq-Sym-cHDP, superimposed with speech words.	51
4.9	Area graph for topic distribution in video No.511 (© <i>Frontline Club</i>) with Seq-Sym-cHDP, superimposed with speech words.	52
4.10	Area graph for topic distribution in video No.984 (© <i>Avelx</i>) with Seq-Sym- cHDP, superimposed with speech words.	53
4.11	Area graph for topic distribution in video No.161 (© <i>The Reality Report</i>) with Seq-CI-HDP, superimposed with speech words.	54
4.12	Area graph for topic distribution in video No.161 (© <i>The Reality Report</i>) with Seq-cHDP, superimposed with speech words.	55
5.1	Graphical model representation of supervised Seq-HDP.	60
5.2	Graphical model representation of Sup-SSC-HDP.	62
5.3	t-SNE 2D embedding of topic frequencies with Sup-SSC-HDP.	68
5.4	t-SNE 2D embedding of topic frequencies with Seq-Sym-cHDP.	69
5.5	Performance of video classification with different λ	70
5.6	Video classification performance of Sup-SSC-HDP and baseline models. . .	71
6.1	A flowchart of framework for video emotion classification.	75
6.2	Graphical representation for stick-breaking construction of Sym-cHDP. . .	79

List of tables

3.1	Summary of Experimental Dataset.	28
3.2	Performance of Genre Classification.	34
4.1	Summary of Experimental Dataset	47
4.2	Performance of Genre Classification.	56
5.1	Summary of Experimental Dataset.	68
5.2	Example Results on Video Classification with Sup-SSC-HDP.	72
6.1	Summary of Experimental Dataset	82
6.2	Unimodel Results with Different Deep Features	83
6.3	Multimodal Results with Different Methods	83

Chapter 1

Introduction

1.1 Background

With the advances in information technology and computer science, research on video data analysis has been attracting more and more focus in recent years. As a matter of fact, such research works are of great value, since video technology is widely used in various areas such as art, education, entertainment, Internet, digital advertising, etc. Nowadays, video data take up more than half of all consumer Internet traffic, and it is forecasted to increase even more in the future. In this background, analyzing enormous video data via machine learning methods are becoming increasingly necessary. For instance, now we can share any video clips through a video-sharing website such as YouTube while attaching several simple labels or detailed descriptions to let viewers easily find clips that interest them. However, an anonymous video file without any labels or descriptions can also be uploaded to such video-sharing website without being sorted into a specific genre. In fact, these anonymous videos need to be learned and categorized properly, since they may contain very useful and important information. However, it is very challenging for researchers to find an appropriate mathematical model to learn videos, since video data have a more complicated structure that generally includes time dependency and multimodal data. Focusing on this problem, this paper will present a number of Bayesian nonparametric topic modeling methods for modeling and analyzing video data, and give a detailed explanation on the advantages of these methods.

In the field of statistics and machine learning, there is a type of statistical model, called Bayesian model, that assumes a prior distribution for the unknown parameters to be estimated via Bayesian inference method. A Bayesian nonparametric model is a Bayesian model on an infinite-dimensional parameter space, where the parameter space is typically chosen as the

set of all possible solutions for a given learning problem [1]. Through Bayesian modeling methods, a number of general machine learning problems can be effectively tackled, such as regression, clustering, classification, latent variable modeling, sequential modeling, etc.

One of the popular instances of Bayesian modeling methods is Bayesian topic modeling method, which can mine latent topics occurring in a collection of documents by using Bayesian inference. Topic models were originally used for solving text mining problems, which assume that documents are mixtures of topics, where a topic is a probability distribution over words. Some statistical techniques, such as variational inference [2, 3] and Markov Chain Monte Carlo (MCMC) [4, 5], can be utilized to infer the set of latent topics that were responsible for generating the words within documents. Generally, Bayesian topic models can be divided into two fundamental types: Bayesian parametric topic model such as latent Dirichlet allocation (LDA) [2, 5], and Bayesian nonparametric topic model such as hierarchical Dirichlet processes (HDP) [4, 3]. The major difference between these two type of topic models is that the number of the topics is required to be manually set beforehand during the inference phase in Bayesian parametric topic models, while the number of the topics can be automatically estimated when using Bayesian nonparametric topic models to infer the latent topics.

Recently, researchers are trying to utilize such Bayesian topic modeling methods to find an appropriate way to analyze video data. However, for the case of video data analysis, the problem becomes more challenging, since the composition of video data is more complex than the one of general text data. More specifically, video data can be deemed as a type of multimodal data that generally contain a sequence of video frames, audio, and speech transcripts. To handle such video data with topic models, there are mainly three challenging problems that need to be tackled: (1) the way to model the time relativity between neighboring video frames, (2) the way to model the multimodal data within a unified topic model, (3) the way to make the model perform a specific application on video data analysis. These problems will be respectively considered and tackled when devising the models in this paper. Compared with original LDA or HDP, the video-oriented topic model is expected to incorporate more modeling mechanisms that may complicate the model with more parameters and variables. Therefore, I prefer using the Bayesian nonparametric topic models for analyzing video data in this paper, which can reduce the difficulty of model tuning, since the number of the topics can be automatically estimated by model itself during the inference phase.

In addition, this paper will focus on two application scenarios of video data analysis: (1) video semantic analysis, (2) video emotion recognition. According to each application scenario, I will extract appropriate and relative video features from videos to let the pro-

posed nonparametric topic model learn their corresponding latent “topics”, so as to achieve corresponding video data analysis. To demonstrate the effectiveness of each proposed nonparametric topic modeling methods on video data analysis, I will devise a series of experiments to assess the proposed model via both qualitative evaluation and quantitative evaluation, and discuss the advantages of using such proposed nonparametric topic modeling methods.

1.2 Related Work

Original Bayesian topic models [2, 4] are generative models for text documents, which specify a set of probabilistic procedures to generate each document. More specifically, topic models generally assume that text words in each document are generated from a mixture distribution of topics, where each topic is represented as a word distribution.

On the basis of this hypothesis, Blei et al. [2] presented a parametric topic model named latent Dirichlet allocation (LDA), where words in a document are considered arising from a particular set of latent topics, where each latent topic is modeled as a multinomial probability distribution over words from some basic vocabulary. A variational inference algorithm [2, 5] was utilized in LDA to infer the latent topics for observed words within documents. The biggest contribution of LDA is that they introduced Dirichlet priors on those multinomial distributions, which effectively simplifies the problem of statistical inference. However, LDA still has a drawback that the number of the topics in LDA is required to be manually set beforehand during the inference phase. Soon after, a nonparametric topic model named hierarchical Dirichlet processes (HDP) [4] was proposed by Teh et al., which models multiple correlated documents with multiple infinite Dirichlet processes [6] that are connected by sharing a mixture of components (topics) among all the documents. HDP rules that the number of topics can be augmented to infinite in the generative process, therefore during the inference phase, the real number of topics is a variable that can be estimated by the inference method.

LDA and HDP can be considered as two most classic and typical topic models. Based on the theory of LDA and HDP, researchers have tried to extend such topic modeling method to varieties of application fields, especially for multimedia data analysis. Wang et al. [7] proposed an unsupervised learning method on the basis of both the LDA and HDP models to detect activities and interactions occurring in the scene of video. However, they ignored the time-dependency issue among neighboring frames within a video according to their method. Hospedales et al. [8] proposed a Markov clustering topic model (MCTM) extended

from LDA for learning scene characteristics, dynamically screening and identifying irregular spatiotemporal patterns within a video. However, only one modality (visual words) of data was considered in their model. Kuettel et al. [9] proposed a cascaded topic model called dependent Dirichlet processes and hidden Markov models (DDP-HMM) to jointly discover spatio-temporal dependencies of moving agents in complex dynamic scenes within a video clip. However, it is very hard to perform multimodal data modelling due to the structure of DDP-HMM. Das et al. [10] proposed a hybrid system that incorporated a low level multimodal latent topic model for initial keyword annotation, a middle level of concept detectors and a high level module, to extract lingual description for a video. However, they did not consider further dependencies and interactions among those modalities. Yang et al. [11] proposed a correspondence Dirichlet compound multinomial LDA (Corr-DCMLDA) to tackle the burstiness problem of the local features for video data mining. However, they did not give a solution for modeling time-related frames within a video as well. For these issues mentioned in above previous works, I provide corresponding solutions in my proposed topic modeling methods in this paper.

In addition, here I also need to introduce some other related works on topic models proposed for other applications, which inspire me a lot on tackling relative problems for this task. For handling the time-related data, Ren et al. [12] proposed a dynamic HDP to deal with music segmentation problem and gene expression problem, where a dynamic relationship between global measure and local measure was modeled in their method. Similarly, Zhang et al. [13] proposed a evolutionary HDP (EvoHDP) model to handle a collection of correlated time-varying text documents, where double time-dependency mechanisms were incorporated in the three-layer EvoHDP. For handling the multimodal data modeling, Blei et al. [14] proposed a correspondence LDA (CorrLDA) model for text-annotated image data, where a unidirectional dependence relationship was predefined between image data and text data. Fukumasu et al. [15] proposed a symmetric correspondence LDA (SymCorrLDA) model for multilingual text analysis, where the dependencies among multiple languages were modeled can controlled by a latent variable. For supervised learning problems on topic models, McAuliffe et al. [16] proposed a supervised latent Dirichlet allocation (sLDA) for predicting movie ratings and document popularity, where a variety of response types can fit in sLDA. Dai et al. [17] proposed a supervised hierarchical Dirichlet processes (sHDP) for the same purpose, where the limited-memory BFGS algorithm was used to find the MAP estimate of regression coefficients.

1.3 Outline

The rest of this paper is organized as follows. Chapter 2 gives a review on the background knowledge of HDP, which is considered to be a base model in this work. Chapter 3 presents a sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) model extended from original HDP model, to analyze semantic information for videos. For the same purpose, Chapter 4 attempts to improve the flexibility of video multimodal data modeling mechanism in Seq-cHDP, then presents a sequential symmetric correspondence hierarchical Dirichlet processes (Seq-Sym-cHDP) model, which can flexibly model the dependency among multimodal data with pivot assignments. Chapter 5 presents a supervised learning version of Seq-Sym-cHDP model for real video semantic analysis application. Chapter 6 develops a Bayesian nonparametric multimodal data modeling framework to mine the latent human emotional events from both deep features and MFCC features extracted from videos, where deep features can be drawn from fully connected layers of convolutional neural networks (CNNs) that is used for deeply learning the images (frames) within each video. Finally, Chapter 7 concludes this work and summarizes overall contributions.

Chapter 2

Preliminary Work

2.1 Background

As mentioned in Chapter 1, original Bayesian topic models [2, 4] are generative models for text documents, which specify a set of probabilistic procedures to generate each document. More specifically, topic models generally assume that text words in each document are generated from a mixture distribution of topics, where each topic is represented as a word distribution. In real application, given a number of the observed documents, I can derive a Bayesian inference method of topic model to learn each latent topic for each given word of documents. Therefore, topic modeling method is usually utilized for discovering and analyzing hidden structures of documents.

In general, topic models can be divided into two fundamental types: Bayesian parametric topic model and Bayesian nonparametric topic model. For Bayesian parametric topic model, the most typical and popular one is latent Dirichlet allocation (LDA) proposed by Blei et al. [2]. In LDA, words in a document are considered arising from a particular set of latent topics, where each latent topic is modeled as a multinomial probability distribution over words from some basic vocabulary. Although the inference of LDA is very easy to be implemented, however, it still has some limitations for real applications since the number of the topics in LDA needs to be predefined beforehand. This may bring much difficulties for model tuning, especially when there are more controlling parameters existing in the model. To tackle this problem, Bayesian nonparametric topic model can be a good option. Based on the theory of Dirichlet process [6], Teh et al. presented a hierarchical Dirichlet processes (HDP) [4] model to mine the latent topics for documents in a Bayesian nonparametric modeling way. In HDP, multiple correlated documents are modeled as multiple infinite Dirichlet processes, which are connected via sharing a mixture of components (topics) among all the documents. Note

that the number of topics is set as infinite in HDP. Therefore, during the inference period of HDP, the real topic counts might be changed for each iteration, and finally converge to an optimized number.

According to the scenario of video data analysis, there are a lot of data features that are required to consider when modeling such video data with topic models. For instance, a video comprises a sequence of video frames where neighboring video frames are time dependent, furthermore, video data are multimodal data that contain various types of data such as image data, audio data and transcript data. To better handle such video data, I need to formulate the topic model with a more complex structure corresponding to the features of video data, which means the modified topic model might incorporate more variables and parameters within its framework. In this case, nonparametric HDP that can automatically estimate the number of the topics, is appropriate to be a base model for modeling video data. I briefly review the generative process and inference method of HDP in the following sections.

2.2 Hierarchical Dirichlet Processes

2.2.1 Generative Process

A Dirichlet process can be deemed as a distribution of a random discrete probability measure. Given a collection of random correlated measures $\{G_j\}$, the HDP connects these random measures probabilistically by using a shared random global measure G_0 , which itself is drawn from a Dirichlet process with base measure H . The graphical model representation for HDP is illustrated in Fig. 2.1. The detailed generative process of HDP is as follows.

- **Step 1:** The global measure G_0 within HDP is drawn from a Dirichlet process:

$$G_0 \sim DP(\gamma, H) \quad (2.1)$$

where $DP(\gamma, H)$ indicates a Dirichlet process with concentration parameter γ and base measure H .

- **Step 2:** Similarly, each local measure G_j is drawn from a Dirichlet process with the global measure G_0 :

$$G_j \sim DP(\alpha_0, G_0) \quad (2.2)$$

where α_0 indicates a concentration parameter for generating G_j .

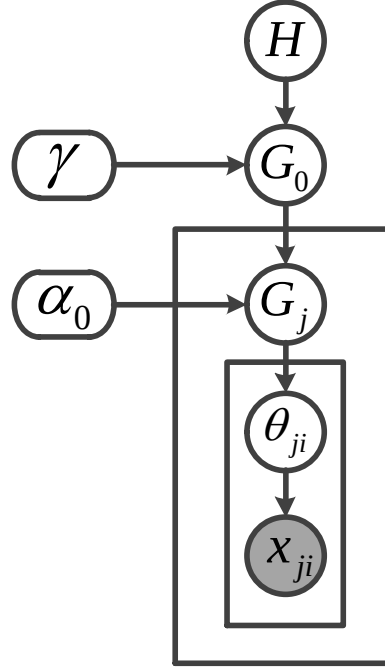


Fig. 2.1 Graphical model representation of HDP.

• **Step 3:** For each j , I define $\{\theta_{ji}\}$ as *i.i.d.* random variables distributed on G_j , then the rest of the generative process is as follows:

$$\begin{aligned}\theta_{ji} &\sim G_j \\ x_{ji} &\sim f(x_{ji}|\theta_{ji})\end{aligned}\tag{2.3}$$

where each θ_{ji} is a factor corresponding to a single observation x_{ji} , and $f(\cdot|\theta_{ji})$ indicates a conditional probability distribution conditioned on θ_{ji} .

To make the model more understandable and applicable, Teh et al. also introduced three analogs from three different perspectives for HDP [4], which are stick-breaking construction, Chinese restaurant process representation, and a representation in terms of a limit of finite mixture models. In this paper, I focus on reviewing the first two methods, described in the following.

2.2.2 Stick-Breaking Construction

Measures extracted from Dirichlet processes can be expressed by using independent sequences of *i.i.d.* random variables. This method is known as a stick-breaking construction [18, 4].

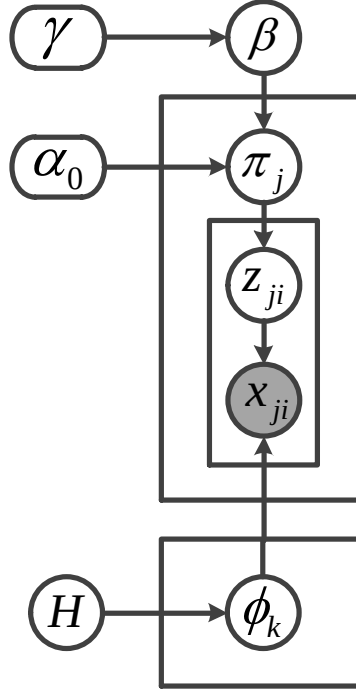


Fig. 2.2 Graphical model representation for stick-breaking construction of Seq-HDP.

The graphical model representation for the stick-breaking construction of HDP is shown in Fig 2.2. Firstly, I define that the components (topics) $\phi = \{\phi_k\}_{k=1}^{\infty}$ are distributed on the base measure H . Since G_0 and G_j have completely support at the points ϕ , I then respectively use component weights β and π_j to construct corresponding measures G_0 and G_j , the process is as below:

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\ G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \end{aligned} \tag{2.4}$$

where each component weight variable consists of an infinite number of corresponding component weights, *i.e.*, $\beta = \{\beta_1, \beta_2, \dots, \beta_k, \dots\}$ and $\pi_j = \{\pi_{j1}, \pi_{j2}, \dots, \pi_{jk}, \dots\}$.

Then, these component weights can be successively drawn with:

$$\begin{aligned} \beta &\sim GEM(\gamma) \\ \pi_j &\sim DP(\alpha_0, \beta) \end{aligned} \tag{2.5}$$

where $GEM()$ stands for a GEM process [19], which is formed by such a process: $\hat{\beta}_k \sim Beta(1, \gamma)$, $\beta_k = \hat{\beta}_k \prod_{i=1}^{k-1} (1 - \hat{\beta}_i)$.

In HDP, each factor θ_{ji} is distributed on G_j , it can take on the value ϕ_k with probability π_{jk} . I define z_{ji} as an indicator variable of ϕ_k , then there is $\theta_{ji} = \phi_{z_{ji}}$. The rest of the generative process is as follows:

$$\begin{aligned} z_{ji} &\sim \pi_j \\ x_{ji} &\sim f(x_{ji} | \phi_{z_{ji}}) \end{aligned} \quad (2.6)$$

2.2.3 Chinese Restaurant Franchise

The Chinese restaurant franchise (CRF) [4] is a metaphor that describes the generative process of topic models via Chinese restaurant process. According to a CRF, there is a restaurant franchise with a shared menu across the restaurants. I assume that each Chinese restaurant can have infinite tables at which customers can sit, each table can serve only one dish for customers, and multiple tables in the same or different restaurants can serve the same dish. When the first customer comes to the restaurant, he or she chooses a table at which to sit and then orders a dish from the global menu or asks the waiter for a new dish not included on the global menu. Then, the next customer comes to the restaurant. He or she can sit at the same table if he or she is willing to eat the same dish as the previous customer. Otherwise, he or she can sit at another table and order any dish he or she wants. Note that the dish can also be a new dish if he or she is not interested in the dishes written on the global menu. If a new dish is ordered, the global menu will be updated and augmented with it.

For a HDP, the documents can be deemed as the restaurants in CRF, and the factors θ_{ji} can be deemed as the customers. I define the $\phi = \{\phi_k\}_{k=1}^K$ distributed on the base measure H as shared global menu of dishes, where K will be incremented later on. I also define variable ψ_{jt} that stands for the specific selection of dish at table t of restaurant j . Then, I depict such a CRF with these notations in Fig. 2.3, where rectangles represent the restaurants, and circles represent the tables of restaurants.

I now compute marginals under a HDP when G_0 and G_j are integrated out. Given $\{\theta_{j1}, \dots, \theta_{j(i-1)}\}$, the θ_{ji} is drawn from:

$$\theta_{ji} | \theta_{j1}, \dots, \theta_{j(i-1)}, \alpha_0, G_0 \sim \sum_{t=1}^{m_j} \frac{n_{jt}}{i-1 + \alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1 + \alpha_0} G_0 \quad (2.7)$$

where n_{jtk} indicates the number of customers who sit at table t of restaurant j eating dish k , and m_{jk} indicates the number of tables in restaurant j serving dish k . All the marginal

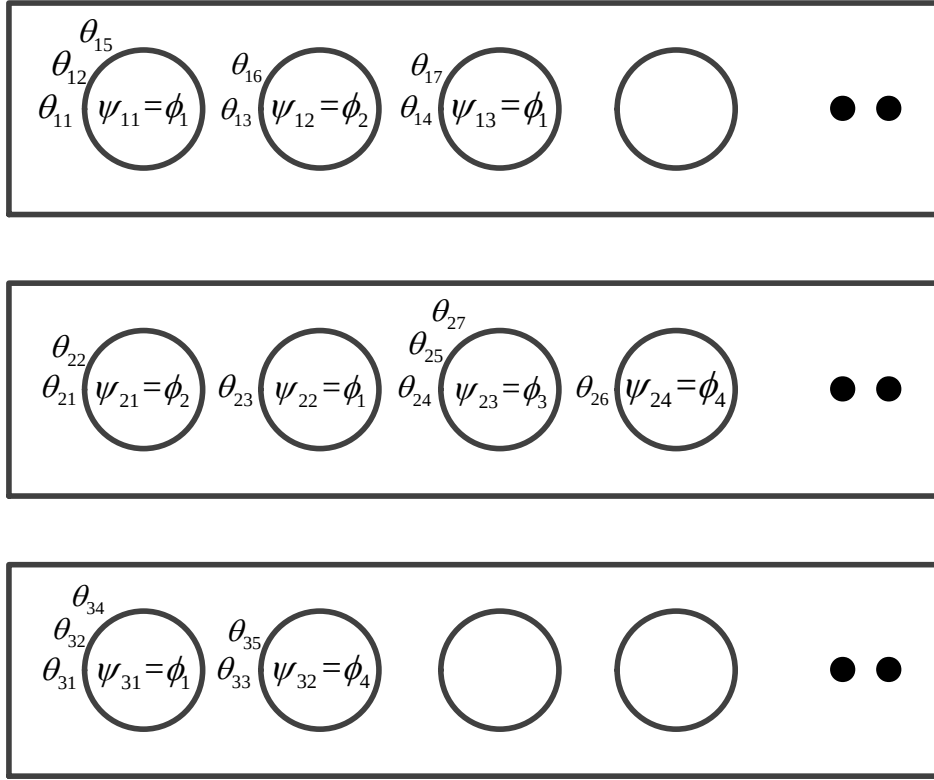


Fig. 2.3 A depiction of a Chinese restaurant franchise.

counts are represented with dots. Note that this process can also be considered as how the customer selecting a table. In this process, the first term is a mixture of the likelihood of existing tables, while the second term is the likelihood of a new-born table. If the second term is selected, I should increment m_j by one.

Next, given $\{\psi_{11}, \psi_{12}, \dots, \psi_{21}, \dots, \psi_{j(i-1)}\}$, the ψ_{ji} is drawn from:

$$\psi_{ji} | \psi_{11}, \psi_{12}, \dots, \psi_{21}, \dots, \psi_{j(i-1)}, \gamma, H \sim \sum_{k=1}^K \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H \quad (2.8)$$

note that this process can also be considered as how the table selecting a dish. In this process, the first term is a mixture of the likelihood of existing dishes, while the second term is the likelihood of a new-born dish. If the second term is selected, I should increment K by one.

2.3 Inference Method for HDP

In general, there are four approaches [4, 20] for inferring the HDP with Markov chain Monte Carlo (MCMC) sampling schemes: (1) a straightforward Gibbs sampler based on the CRF, (2) an augmented representation involving both the CRF and the posterior for G_0 , (3) a variation on the second sampling scheme with streamlined bookkeeping, and (4) a posterior representation sampler from the perspective of the stick-breaking construction. In this section, I focus on introducing the fourth one.

According to the scheme of the posterior representation sampler, the global measure G_0 can be formulated by:

$$G_0 = \sum_{k=1}^K \beta_k \delta_{\phi_k} + \beta_u G_u, \quad G_u \sim DP(\gamma, H) \quad (2.9)$$

and the component weights β can be sampled from:

$$\beta = (\beta_1, \dots, \beta_K, \beta_u) \sim Dir(m_{.1}, \dots, m_{.K}, \gamma) \quad (2.10)$$

where *Dir* indicates Dirichlet distribution. In this procedure, I reformulate the original infinite vector β with finite K components and a promising component u . As mentioned in CRF, m_{jk} indicates the number of tables in restaurant j serving dish k , here its marginal form is $m_{.k}$ that affects the estimation of β_k . The sampled value of the promising component weight β_u is influenced by the concentration parameter γ .

Similarly, each local measure G_j can be formulated by:

$$G_j = \sum_{k=1}^K \pi_{jk} \delta_{\phi_k} + \pi_{ju} G_u, \quad G_u \sim DP(\gamma, H) \quad (2.11)$$

and the component weights π_j can be sampled from:

$$\pi_j = (\pi_{j1}, \dots, \pi_{jK}, \pi_{ju}) \sim Dir(\alpha_0 \beta_1 + n_{j1}, \dots, \alpha_0 \beta_K + n_{jK}, \alpha_0 \beta_u) \quad (2.12)$$

where I reformulate the original infinite vector π_j with finite K components and a promising component u . n_{jk} indicates the number of samples that are assigned with topic k in document j . Note that both β_k and n_{jk} affect the estimation of π_{jk} , and the sampled value of the promising component weight π_{ju} is influenced by $\alpha_0 \beta_u$.

Algorithm 1: Gibbs Sampling Method for HDP

```

Initialization; Sample  $\beta$ ;
for ( $j = 1; j \leq J; j++$ ) do
    Sample  $\pi_j$ ;
    for ( $i = 1; i \leq N_j; i++$ ) do
        Sample  $z_{ji}$ ;
    end
    Compute each  $n_{jk}$ ;
    Sample each  $m_{jk}$ ;
end
Compute each  $m_{.k}$ ;

```

Then, for each sample x_{ji} , I derive a conditional likelihood to sample the topic z_{ji} for each x_{ji} :

$$\begin{aligned}
 p(z_{ji} = k | x_{ji}, \dots) &\propto p(z_{ji} = k | \pi_j) p(x_{ji} | z_{ji} = k, \dots) \\
 &= \begin{cases} \pi_{jk} f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ is used} \\ \pi_{ju} f_{k^{new}}^{-x_{ji}}(x_{ji}) & \text{if } k \text{ is newborn} \end{cases}
 \end{aligned} \tag{2.13}$$

where $f_k^{-x_{ji}}(x_{ji})$ and $f_{k^{new}}^{-x_{ji}}(x_{ji})$ are the sample probabilities conditioned on the topic k , and the superscript $-x_{ji}$ denotes doing the computation excluding the current sample. Note that both of $f_k^{-x_{ji}}(x_{ji})$ and $f_{k^{new}}^{-x_{ji}}(x_{ji})$ can be defined in a variety of ways according to different applications.

The n_{jk} can be computed by counting all the z_{ji} . However, I still need to find a way to sample the counts m_{jk} . To simplify this problem, I can use a Bernoulli trial method proposed by Porteous [21] to sample it. I give the estimation of m_{jk} as below:

$$p(m_{jkr} = 1) = \frac{\alpha_0 \beta_k}{\alpha_0 \beta_k + r - 1} \tag{2.14}$$

where r is the index of the customers and $r \in [1, n_{jk}]$, $p(m_{jkr} = 1)$ indicates the probability of that the r -th customer chooses a new table at which to eat dish k in the restaurant j . Hence, every m_{jkr} can be sampled via this approach, then compute the m_{jk} via $m_{jk} = \sum_r m_{jkr}$.

Then, I employ a Gibbs sampling method to implement such posterior inference of HDP. Its pseudo-code is shown in Algorithm 1.

In this method, we need to pay attention because the number of topics k may increase if a newborn topic is selected during sampling z_{ji} . For this case, all the component weights need to update their atoms to instantiate this newborn topic in the sampler, whose procedures are

expressed with:

$$\begin{aligned}
h|\gamma &\sim \text{Beta}(\gamma, 1) \\
(\beta_u^{new}, \beta_{(K+1)}^{new}) &= (\beta_u h, \beta_u (1 - h)) \\
h_j|\alpha_0, \beta_u, h &\sim \text{Beta}(\alpha_0 \beta_u h, \alpha_0 \beta_u (1 - h)) \\
(\pi_{ju}^{new}, \pi_{j(K+1)}^{new}) &= (\pi_{ju} h_j, \pi_{ju} (1 - h_j))
\end{aligned} \tag{2.15}$$

where $\text{Beta}()$ represents beta distribution. After this update, the total number of the topics K will be increased by one. In addition, I just show only one iteration of the whole Gibbs sampling task in Algorithm 1. As a matter of fact, it might operate thousands of iterations to let all the variables reach convergence depending on different scenarios of applications.

Chapter 3

Sequential Correspondence Hierarchical Dirichlet Processes for Video Semantic Analysis

3.1 Overview

In this chapter, I present a Bayesian nonparametric topic model named sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) for video semantic analysis.

As I mentioned, straightforwardly applying HDP to deal with video data is not a good solution, since video data have a more complicated structure than text data, *i.e.*, a video can be deemed as a type of multimodal data, which generally contain image data and speech data from the perspective of video semantic analysis. In this work, for a specific video, I consider that a sequence of video frames represent image data, and speech transcripts represent speech data.

To deal with such video data, I extend the original HDP, and develop a more appropriate Bayesian nonparametric topic model, *i.e.*, Seq-cHDP, which involves two core aspects: one is the time-dependency mechanism that shows the time relativity between neighboring video frames based on a Markovian assumption, and the other is the data correspondence mechanism that provides two corresponding generative processes for the multimodal data within the model. From the perspective of Bayesian nonparametric topic models, the observations are discrete samples (words). Hence, the preprocessing for videos is necessary, where image data and speech data are respectively converted into a number of corresponding visual words and speech words. Since the data correspondence mechanism defines a unidirectional de-

pendence relationship among multimodal data, this work assumes image data as a pivot data type (modality), and the latent topics of non-pivot speech data are sampled from the topic distribution over pivot image data. The generative process of Seq-cHDP is reformulated with stick-breaking construction. Then, an inference method is developed based on posterior representation sampler, and implemented by cascaded Gibbs sampling. In experimentations, Seq-cHDP is assessed by topic trend estimation and genre classification.

The rest of this chapter is organized as follows. Section 3.2 presents a sequential HDP (Seq-HDP) to handle the time-dependency issue about neighboring frames within a video. Based on Seq-HDP, Section 3.3 incorporates a data correspondence mechanism to model the multimodal data in videos. Section 3.4 develops an inference method based on posterior representation sampler to infer the Seq-cHDP. Finally, experimental evaluations and concluding remarks are respectively presented in Section 3.5 and Section 3.6.

3.2 Sequential HDP

In this section, I focus on developing a unimodal sequential HDP (Seq-HDP) model to handle the time-dependency issue about neighboring video frames. Seq-HDP extended from HDP has a three-layer hierarchy in accordance with the structure of video data. I firstly describe the generative process of Seq-HDP in Section 3.2.1. Then, a stick-breaking construction for Seq-HDP is introduced in Section 3.2.2.

3.2.1 Generative Process

Inspired by the discussions on modeling the time-varying activities in the previous studies [12, 13], here I present a simple three-layer HDP model called Seq-HDP to show the time dependencies among neighboring frames within each video. I depict the graphical model representation of Seq-HDP in Fig. 3.1. In this three-layer Seq-HDP model, a measure G located in the top of the model structure represents an overall measure of Seq-HDP, which is generated from a base measure H and simultaneously shared by all the HDPs. Similarly, I define G_0^f as the global measure reflecting the f -th video on the second layer of Seq-HDP. Then on the third layer, G_j^f denotes the local measure corresponding to the j -th frame within the f -th video. As a matter of fact, these local measures are not independent, since video frames within the same video are sequential and time correlated. Therefore, as shown in the graphical model representation of Seq-HDP, I establish a transfer link based on a simple

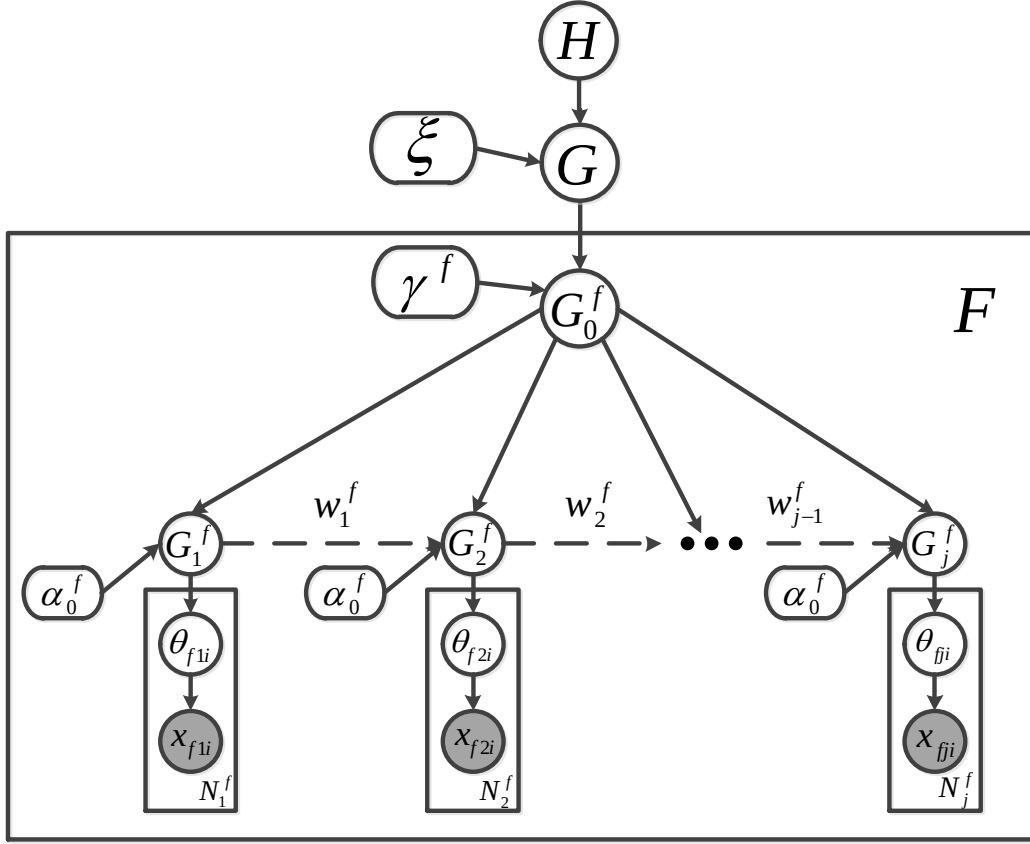


Fig. 3.1 Graphical model representation of Seq-HDP.

Markovian assumption between two adjacent local measures G_{j-1}^f and G_j^f in Seq-HDP with a transfer weight w_j^f . The entire generative process of Seq-HDP is described below.

- **Step 1:** The overall measure G is drawn from a Dirichlet process:

$$G \sim DP(\xi, H) \quad (3.1)$$

where $DP()$ represents a Dirichlet process, ξ is a concentration parameter for generating G .

- **Step 2:** The global measure G_0^f for each video is drawn from:

$$G_0^f \sim DP(\gamma^f, G) \quad (3.2)$$

where f is the index of the video, and γ^f is a concentration parameter for generating G_0^f .

- **Step 3:** According to the time-dependency mechanism, both the upper global measure G_0^f and previous local measure G_{j-1}^f affect the generation of G_j^f . The local measure G_j^f is drawn

from:

$$G_j^f \sim DP(\alpha_0^f, w_{j-1}^f G_{j-1}^f + (1 - w_{j-1}^f) G_0^f) \quad (3.3)$$

where α_0^f is a concentration parameter for generating G_j^f , and w_{j-1}^f is the transfer weight.

• **Step 4:** For each frame j within video j , the generative process for each random variable θ_{fji} and each observation (data sample) x_{fji} is as follows:

$$\begin{aligned} \theta_{fji} &\sim G_j^f \\ x_{fji} &\sim f(x_{fji} | \theta_{fji}) \end{aligned} \quad (3.4)$$

where each θ_{fji} drawn from local measure G_j^f is a factor corresponding to a single data sample x_{fji} , the data sample x_{fji} is sampled from a conditional probability distribution parameterized with θ_{fji} . Note that i is the index of the data sample, and $f(\cdot|\cdot)$ represents a conditional probability distribution.

3.2.2 Stick-Breaking Construction

As mentioned in Section 2.2.2, the stick-breaking construction is employed as a type of analog for nonparametric topic models, which reformulates the nonparametric topic model by using independent sequences of *i.i.d.* random variables.

The graphical model representation for the stick-breaking construction of Seq-HDP is illustrated in Fig. 3.2. Here, I also define that the components (topics) $\phi = \{\phi_k\}_{k=1}^{\infty}$ are distributed on the base measure H . The infinite component weights can be used for constructing each probability measure. The overall measure G can be reformulated by:

$$\begin{aligned} G &= \sum_{k=1}^{\infty} v_k \delta_{\phi_k} \\ v &\sim GEM(\xi) \end{aligned} \quad (3.5)$$

where infinite vector v indicates a collection of overall component weights corresponding to G , *i.e.*, $v = \{v_1, v_2, \dots, v_k, \dots\}$, and the overall component weights v are obtained via a GEM distribution [19] parameterized with ξ , which is specified in Section 2.2.2 of Chapter 2.

Similarly, the global measure G_0^f is reformulated by:

$$\begin{aligned} G_0^f &= \sum_{k=1}^{\infty} \beta_k^f \delta_{\phi_k} \\ \beta^f &\sim DP(\gamma^f, v) \end{aligned} \quad (3.6)$$

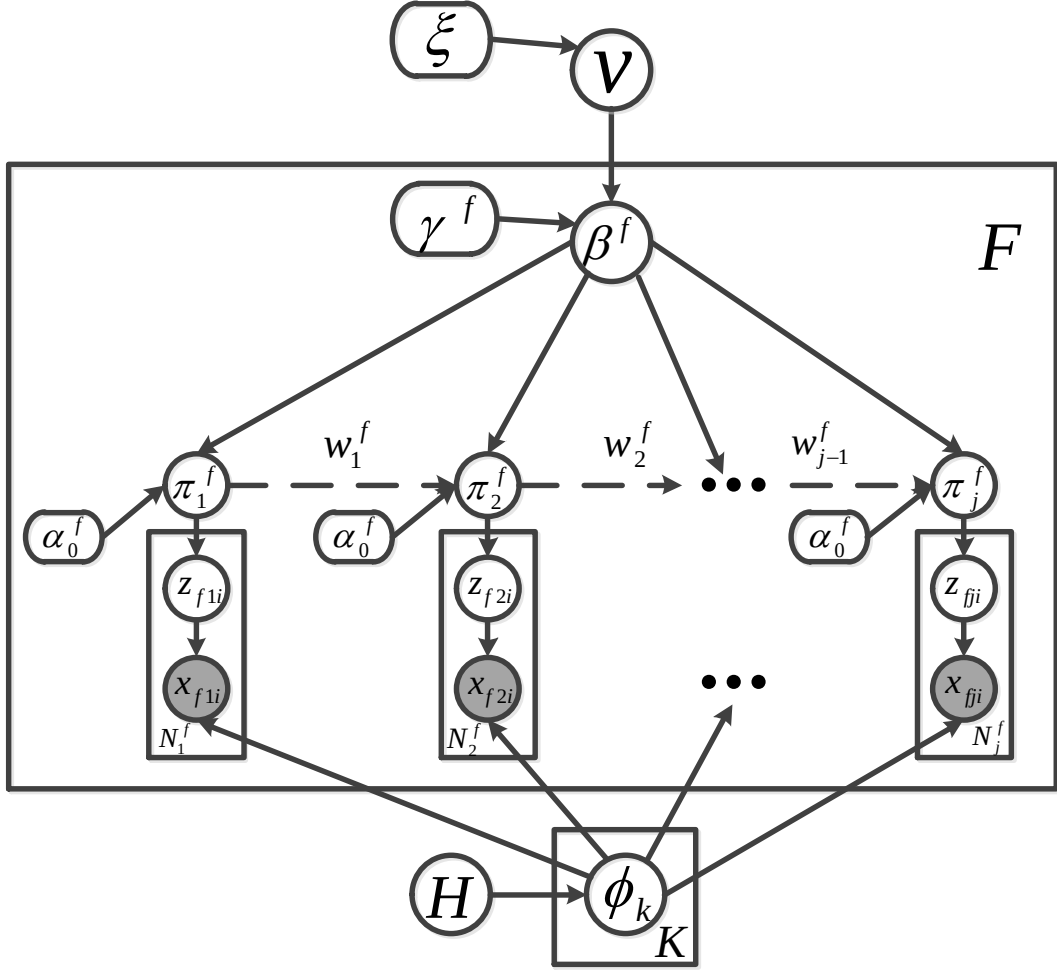


Fig. 3.2 Graphical model representation for stick-breaking construction of Seq-HDP.

where infinite vector β^f indicates a collection of global component weights corresponding to G_0^f and $\beta^f = \{\beta_1^f, \beta_2^f, \dots, \beta_k^f, \dots\}$. The global component weights β^f are drawn from a Dirichlet process parameterized with γ^f and v .

In the last layer, the local measure G_j^f is reformulated by:

$$G_j^f = \sum_{k=1}^{\infty} \pi_{jk}^f \delta_{\phi_k} \quad (3.7)$$

$$\pi_j^f \sim DP(\alpha_0^f, \pi_j^f)$$

where infinite vector π_j^f indicates a collection of local component weights corresponding to G_j^f and $\pi_j^f = \{\pi_{j1}^f, \pi_{j2}^f, \dots, \pi_{jk}^f, \dots\}$. The local component weights π_j^f are drawn from a Dirichlet process parameterized with α_0^f and π_j^f . Because of the time-dependency mechanism, the

π_j^f is formed as follows.

$$\pi_j^f = w_{j-1}^f \pi_{j-1}^f + (1 - w_{j-1}^f) \beta^f \quad (3.8)$$

Then, I use z_{fji} substitutes the factor θ_{fji} in Seq-HDP, which can be deemed as indicator variable of component ϕ_k . The rest of the generative process is as follows:

$$\begin{aligned} z_{fji} &\sim \pi_j^f \\ x_{fji} &\sim f(x_{fji} | \phi_{z_{fji}}) \end{aligned} \quad (3.9)$$

3.3 Correspondence Method for Multimodal Data Modeling

As mentioned in Section 3.1, video data can be deemed as a type of multimodal data, which generally include image data and speech data from the perspective of video semantic analysis. In this section, I incorporate a correspondence method inspired by a previous work [14] into the Seq-HDP model to handle such multimodal data. The new model is named sequential correspondence hierarchical Dirichlet processes (Seq-cHDP).

According to this correspondence mechanism, a unidirectional dependence relationship is predefined among multimodal data. More specifically, the correspondence mechanism predefines a pivot data type (modality), and then makes other non-pivot data types (modalities) generate their latent topics based on the topic distribution over pivot data. In the scenario of video data analysis with Seq-cHDP, I assume that image data is the pivot data, and image data and speech data are discretized with visual words and speech words, respectively.

I show a graphical model representation for the stick-breaking construction of Seq-cHDP in Fig. 3.3, where the components shared in the model indicate latent topics of the words. The data sample generative process is clearly separated into two different blocks, which respectively represent the generative methods of visual words and speech words. The generative process of Seq-cHDP is the same as that of Seq-HDP until generating the parameter π_j^f . Then, the rest of the generative process of Seq-cHDP is as follows.

- **Step 1:** For each frame j , a topic z_{fji} for a visual word is drawn from π_j^f . Then, conditioned on the sampled topic z_{fji} , a specific visual word x_{fji} is drawn from $f(x_{fji} | \phi_k, k = z_{fji})$.
- **Step 2:** In the same frame, a topic $\hat{z}_{fji}$ for a speech word is drawn from $Uniform(z_{fj1}, z_{fj2}, \dots, z_{fjN_{fj}})$. Then, conditioned on the sampled topic $\hat{z}_{fji}$, a specific speech word $\hat{x}_{fji}$ is drawn from $f(\hat{x}_{fji} | \hat{\phi}_k, k = \hat{z}_{fji})$.

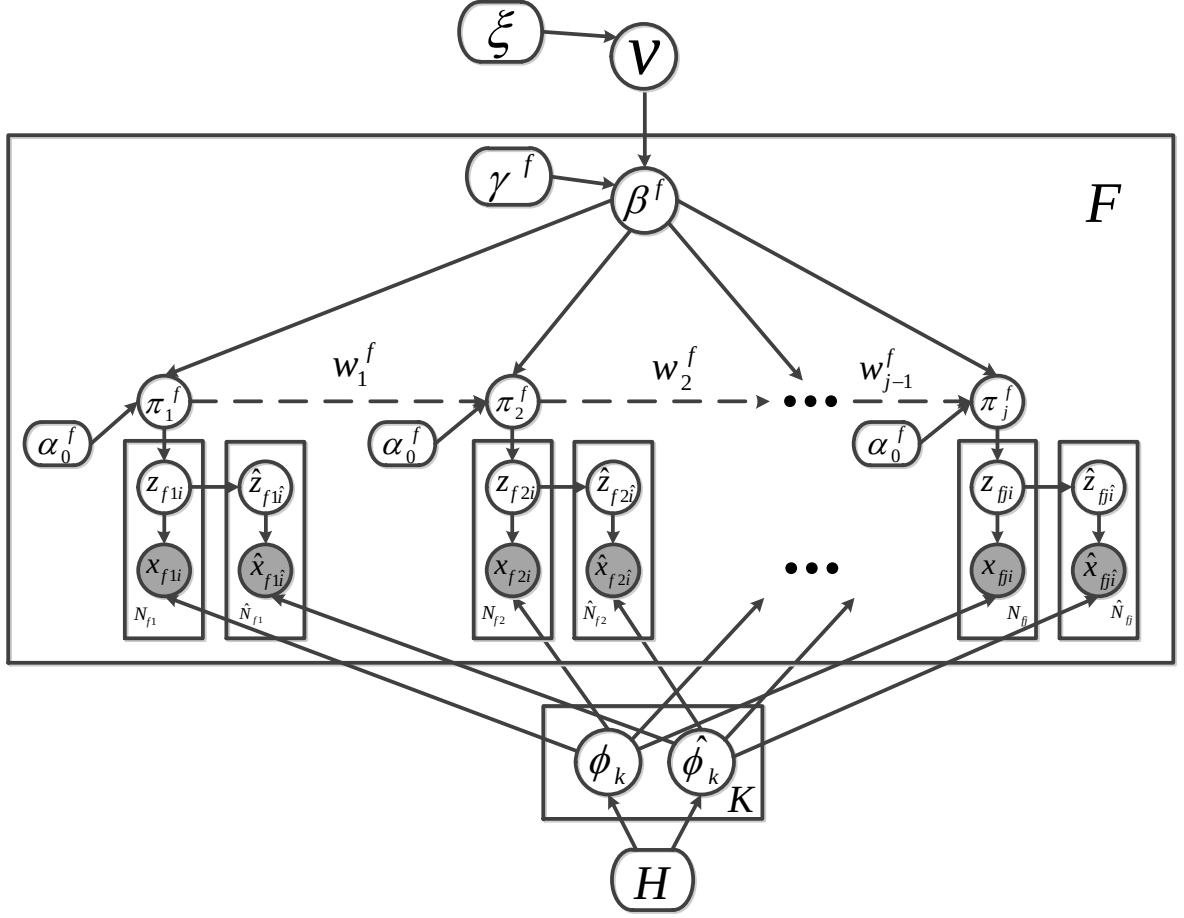


Fig. 3.3 Graphical model representation for stick-breaking construction of Seq-cHDP.

Here, $Uniform()$ denotes a uniform distribution. N_{fj} and $\hat{N}_{fj}$ respectively indicate the number of visual words and speech words in frame j of video f . $f(\cdot|\cdot)$ indicates a conditional distribution. As a matter of fact, there are two different topic collections $\{\phi_k\}_{k=1}^{\infty}$ and $\{\hat{\phi}_k\}_{k=1}^{\infty}$ drawn from the same base measure H that respectively belong to visual words and speech words. However, they share the same topic space. Note that the generative process of the topics for speech words is uncorrelated with topic weights π_j^f , and that it only depends on the sampled topics for the pivot visual words in the same frame. This is also recognized as a key character for the correspondence mechanism.

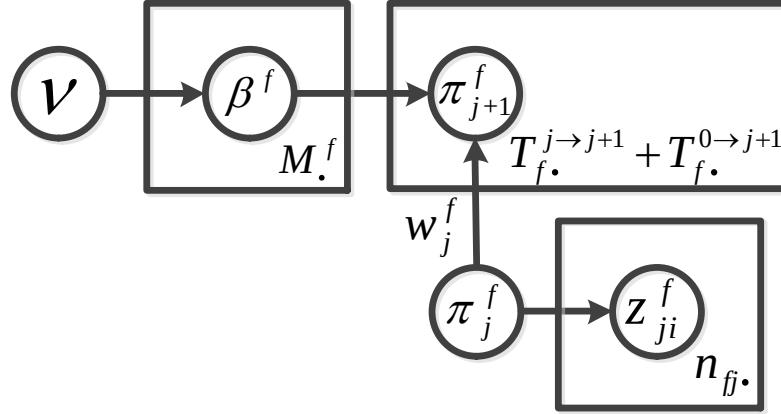


Fig. 3.4 Generation mechanism of stick-breaking construction for Seq-cHDP associated with the extracted counts of factors in CRF.

3.4 Inference Method

An inference method for Seq-cHDP will be introduced in this section. I derive posterior representation sampler to infer the Seq-cHDP, and implement it with a cascaded Gibbs sampling method.

3.4.1 Factors in Chinese Restaurant franchise

Referring to the introduction of original HDP [4], the model of Chinese restaurant franchise (CRF) is derived to reformulate the generative process of HDP from another perspective. According to the hypothesis of CRF in HDP, it assumes that a number of Chinese restaurants serve customers dishes from a shared menu, where a customer represents a single word, a dish represents a latent topic, and the shared menu represents the collection of topics. The detailed generative process for CRF is described in Section 2.2.3 of Chapter 2.

However, for the scenario of Seq-cHDP, the reformulation with CRF becomes more complex, since the Seq-cHDP has a three-layer hierarchy. Referring to the theory in the previous work [13], a new factor metatable (table of the tables) is introduced for integrating out the overall measure G in a three-layer HDP. I also notice that the extracted counts of these factors (*i.e.*, customer, table and metatable) in CRF are correlated with the component weights v , β^f and π_j^f in stick-breaking construction, since these counts are necessary for updating the corresponding component weights in the posterior representation sampler.

In the following, I will describe a brief generation mechanism of stick-breaking construction for Seq-cHDP associated with the extracted counts of factors in CRF, as shown in

Fig. 3.4. M_k^f indicates the number of the metatables in the area (video) f serving dish k that is drawn from the overall menu $\mathbf{v}$. $T_{(j+1)k}^f$ indicates the number of the tables in the restaurant $j+1$ of area (video) f serving dish k . Because of the time dependency, this dish can be selected from the previous local menu π_j^f or the global menu β^f . Hence, $T_{(j+1)k}^f$ is divided into two types with $T_{(j+1)k}^f = T_{fk}^{j \rightarrow j+1} + T_{fk}^{0 \rightarrow j+1}$, where $T_{fk}^{j \rightarrow j+1}$ indicates the number of tables whose dishes are drawn from π_j^f , and $T_{fk}^{0 \rightarrow j+1}$ indicates the number of tables whose dishes are drawn from β^f . Then, n_{fjk} indicates the number of the customers eating dish k in the restaurant j of area (video) f , and the ordered dishes are drawn from the local menu π_j^f . Note that all the customers (words) in n_{fjk} belong to the modality of visual words, since I predefine that the modality of visual words is the pivot modality in Seq-cHDP, whose topics are straightforwardly drawn from π_j^f . In addition, all the topic index k on the counts of factors are marginalized with dots in Fig. 3.4. The sampling method for the counts of these factors in CRF will be specified in Section 3.4.2.

3.4.2 Posterior Representation Sampler

In the following, I will describe how to sample the variables mentioned in the stick-breaking construction for Seq-cHDP.

According to the scheme of the posterior representation sampler for HDP that I described in Section 2.3, the overall measure G can be reformulated by:

$$G = \sum_{k=1}^K v_k \delta_{\phi_k} + v_u G_u, \quad G_u \sim DP(\xi, H) \quad (3.10)$$

and the overall component weights $\mathbf{v}$ can be sampled from:

$$\mathbf{v} = (v_1, \dots, v_K, v_u) \sim Dir(M_1, \dots, M_K, \xi) \quad (3.11)$$

where Dir denotes Dirichlet distribution. As described in Section 3.4.1, M_k is the marginal form of M_k^f , hence $M_k = \sum_f M_k^f$. In this procedure, the original infinite vector $\mathbf{v}$ has an augmentable form filled with finite K components and a promising component u .

Similarly, the global measure G_0^f can be reformulated by:

$$G_0^f = \sum_{k=1}^K \beta_k^f \delta_{\phi_k} + \beta_u^f G_u, \quad G_u \sim DP(\xi, H) \quad (3.12)$$

and the global component weights β^f can be sampled from:

$$\beta^f = (\beta_1^f, \dots, \beta_K^f, \beta_u^f) \sim \text{Dir}(\tilde{\beta}_1^f, \dots, \tilde{\beta}_K^f, \tilde{\beta}_u^f) \quad (3.13)$$

where the estimator $\tilde{\beta}_k^f$ and $\tilde{\beta}_u^f$ are formed with:

$$\begin{aligned} \tilde{\beta}_k^f &= \gamma^f v_k + T_{fk}^{0 \rightarrow \cdot} \\ \tilde{\beta}_u^f &= \gamma^f v_u \end{aligned} \quad (3.14)$$

According to the CRF of Seq-cHDP, only some of the tables $T_{fk}^{0 \rightarrow \cdot}$ assigned with dish k among all the restaurants are associated with β_k^f . Here, $T_{fk}^{0 \rightarrow \cdot} = \sum_j T_{fk}^{0 \rightarrow j}$.

Then, the local measure G_j^f can be reformulated by:

$$G_j^f = \sum_{k=1}^K \pi_{jk}^f \delta_{\phi_k} + \pi_{ju}^f G_u, \quad G_u \sim DP(\xi, H) \quad (3.15)$$

and the local component weights π_j^f can be sampled from:

$$\pi_j^f = (\pi_{j1}^f, \dots, \pi_{jK}^f, \pi_{ju}^f) \sim \text{Dir}(\tilde{\pi}_{j1}^f, \dots, \tilde{\pi}_{jK}^f, \tilde{\pi}_{ju}^f) \quad (3.16)$$

referring to Fig. 3.4, the local component weights π_j^f has two generating paths that respectively generate n_{jf} and $T_f^{j \rightarrow j+1}$. Hence, the estimator $\tilde{\pi}_{jk}^f$ and $\tilde{\pi}_{ju}^f$ are formed with:

$$\begin{aligned} \tilde{\pi}_{jk}^f &= \alpha_0^f w_{j-1}^f \pi_{(j-1)k}^f + \alpha_0^f (1 - w_{j-1}^f) \beta_k^f + n_{jk}^f + T_{fk}^{j \rightarrow j+1} \\ \tilde{\pi}_{ju}^f &= \alpha_0^f w_{j-1}^f \pi_{(j-1)u}^f + \alpha_0^f (1 - w_{j-1}^f) \beta_u^f \end{aligned} \quad (3.17)$$

where the weight w_j^f is a controlling parameter that ranges from 0 to 1.

For the concentration parameters ξ , γ^f and α_0^f , I use a gamma prior to update them:

$$\begin{aligned} \xi &\sim \text{Ga}(a_\xi, b_\xi) \\ \gamma^f &\sim \text{Ga}(a_\gamma, b_\gamma) \\ \alpha_0^f &\sim \text{Ga}(a_{\alpha_0}, b_{\alpha_0}) \end{aligned} \quad (3.18)$$

where a represents the prior shaping parameter, while b represents the prior scaling parameter.

As mentioned in Section 3.4.1, the table counts T_{jk}^f and the metatable counts M_k^f are very necessary for updating the corresponding component weights in the posterior representation

sampler. Here, I also utilize the Bernoulli trial method [21] to sample T_{jk}^f and M_k^f . The estimation of T_{jk}^f is computed via:

$$p(T_{jkr}^f = 1) = \frac{\alpha_0^f [w_{j-1}^f \pi_{(j-1)k}^f + (1 - w_{j-1}^f) \beta_k^f]}{\alpha_0^f [w_{j-1}^f \pi_{(j-1)k}^f + (1 - w_{j-1}^f) \beta_k^f] + r - 1} \quad (3.19)$$

where r indicates the index of the customers and $r \in [1, n_{jk}^f]$, $p(T_{jkr}^f = 1)$ indicates the probability of that the customer r selects a new table at which to eat dish k in the restaurant j of the area f . Then, T_{jk}^f can be obtained by summing up all the T_{jkr}^f , i.e., $T_{jk}^f = \sum_r T_{jkr}^f$.

Similarly, the estimation of M_k^f is computed via:

$$p(M_{kt}^f = 1) = \frac{\gamma^f v_k}{\gamma^f v_k + t - 1} \quad (3.20)$$

where t indicates the index of the tables and $t \in [1, T_k^f]$. Then, the M_k can be computed by $M_k = \sum_f \sum_t M_{kt}^f$ after all the M_{kt}^f have been sampled.

Since the $T_{(j+1)k}^f$ comprises $T_{fk}^{j \rightarrow j+1}$ and $T_{fk}^{0 \rightarrow j+1}$, which can be sampled with a simple multinomial distribution:

$$(T_{fk}^{j \rightarrow j+1}, T_{fk}^{0 \rightarrow j+1}) \sim \text{Multinomial}(T_{(j+1)k}^f, [p, 1 - p]) \quad (3.21)$$

where p indicates a probability for allocating tables that is formulated with:

$$p = \frac{w_j^f \pi_{jk}^f}{(1 - w_j^f) \beta_k^f + w_j^f \pi_{jk}^f} \quad (3.22)$$

According to the correspondence mechanism of Seq-cHDP, I predefine visual words as the pivot modality. Then, the conditional likelihood of topic z_{fji} with respect to visual words is expressed with:

$$\begin{aligned} p(z_{fji} = k | x_{fji}, \dots) &\propto p(z_{fji} = k | \pi_j^f) p(x_{fji} | z_{fji} = k, \dots) \\ &= \begin{cases} \pi_{jk}^f f_k^{-x_{fji}}(x_{fji}) & \text{if } k \text{ is used} \\ \pi_{ju}^f f_{k^{new}}^{-x_{fji}}(x_{fji}) & \text{if } k \text{ is newborn} \end{cases} \end{aligned} \quad (3.23)$$

where $f_k^{-x_{fji}}(x_{fji})$ and $f_{k^{new}}^{-x_{fji}}(x_{fji})$ are the probabilities of visual words conditioned on the topic k . The superscript $-x_{fji}$ represents doing the computation excluding the current sample. According to the theory of HDP-LDA, if the topic k has been previously used, the $f_k^{-x_{fji}}(x_{fji})$

is expressed with:

$$\begin{aligned} f_k^{-x_{fji}}(x_{fji} = v) &= \int f(x_{fji} = v | \phi_k) p(\phi_k | X_k^{-f,ji}, H) d\phi_k \\ &= \frac{n_{kv}^{-fji} + \tau}{n_{k\cdot}^{-fji} + V\tau} \end{aligned} \quad (3.24)$$

and if k is a newborn topic, the $f_{k^{new}}^{-x_{fji}^f}(x_{fji}^f)$ can be expressed with:

$$f_{k^{new}}^{-x_{fji}^f}(x_{fji}^f = v) = \int f(x_{fji}^f = v | \phi_k) p(\phi_k | H) d\phi_k = \frac{1}{V} \quad (3.25)$$

where v indicates the index of visual words in terms of the vocabulary, $X_k^{-f,ji}$ indicates a set of all the visual words assigned to topic k except for x_{fji} , n_{kv}^{-fji} indicates the counts of the visual word v assigned to topic k except for the current one and $n_{k\cdot}^{-fji} = \sum_v n_{kv}^{-fji}$, and V is the total number of visual words in the vocabulary. In addition, τ is a controlling parameter and I define $H = Dir(\tau)$.

Then, conditioned on the sampled topic collection of visual words Z_{fj} , the conditional likelihood of topic $\hat{z}_{fji}$ with respect to speech words is expressed with:

$$\begin{aligned} p(\hat{z}_{fji} = k | \hat{x}_{fji}, \dots) &\propto p(\hat{z}_{fji} = k | Z_{fj}) p(\hat{x}_{fji} | \hat{z}_{fji} = k, \dots) \\ &= \frac{n_{fjk}}{n_{fj\cdot}} \hat{f}_k^{-\hat{x}_{fji}}(\hat{x}_{fji}) \end{aligned} \quad (3.26)$$

and the $\hat{f}_k^{-\hat{x}_{fji}}(\hat{x}_{fji})$ is expressed with:

$$\begin{aligned} \hat{f}_k^{-\hat{x}_{fji}}(\hat{x}_{fji} = \hat{v}) &= \int \hat{f}(\hat{x}_{fji} = \hat{v} | \hat{\phi}_k) p(\hat{\phi}_k | \hat{X}_k^{-f,j\hat{i}}, H) d\hat{\phi}_k \\ &= \frac{\hat{n}_{k\hat{v}}^{-fj\hat{i}} + \tau}{\hat{n}_{k\cdot}^{-fj\hat{i}} + \hat{V}\tau} \end{aligned} \quad (3.27)$$

where $\hat{v}$ indicates the index of speech words in terms of the vocabulary, and $n_{fj\cdot}$ is the marginal form for the counts of visual words as $n_{fj\cdot} = \sum_k n_{fjk}$. Similarly, $\hat{n}_{kv}^{-fj\hat{i}}$ indicates the counts of the speech word v assigned to topic k except for the current one and $\hat{n}_{k\cdot}^{-fj\hat{i}} = \sum_{\hat{v}} \hat{n}_{k\hat{v}}^{-fj\hat{i}}$, $\hat{V}$ is the total number of speech words in the vocabulary. Note that there is no newborn k sampled in this process, since all the sampled topics with respect to speech words conform to the topic space of the sampled topic collection of visual words Z_{fj} .

Algorithm 2: Cascaded Gibbs Sampling for Inferring Seq-cHDP

```

Initialization;
for ( $f = 1; f \leq F; f++$ ) do
    for ( $j = J^f; j \geq 1; j--$ ) do
        Sample topics for visual words  $Z_{fj}$ ;
        Sample topics for speech words  $\hat{Z}_{fj}$ ;
        Sample table counts  $T_{jk}^f, T_{fk}^{j-1 \rightarrow j}$  and  $T_{fk}^{0 \rightarrow j}$ ;
    end
    Sample metatable counts  $M_k^f$ ;
end
Sample concentration parameters  $\xi, \gamma^f$  and  $\alpha_0^f$ ;
Sample overall component weights  $v$ ;
for ( $f = 1; f \leq F; f++$ ) do
    Sample global component weights  $\beta^f$ ;
    for ( $j = 1; j \leq J^f; j++$ ) do
        Sample local component weights  $\pi_j^f$ ;
    end
end

```

3.4.3 Cascaded Gibbs Sampling

To implement the inference method for Seq-cHDP, I also employ the Gibbs sampling approach based on posterior representation sampler. Different from the Gibbs sampling approach used in original HDP, the implementation for inferring Seq-cHDP is mainly separated into two individual sampling procedures due to time-dependency mechanism: the first step is to sample the latent topics, table counts and metatable counts, the second step is to sample the three different component weights. This sampling strategy is also called cascaded Gibbs sampling. The pseudo-code for this cascaded Gibbs sampling method is shown in Algorithm 2.

This algorithm only show one iteration of the whole Gibbs sampling task. J^f indicates the number of the frames in video f , while F indicates the total number of the videos. Note that the number of topics k may increase if a newborn topic is selected during the sampling task. For this situation, all the component weights need to update their atoms to instantiate this new topic in the sampler. This updating method has been already specified in Section 2.3 of Chapter 2.

Table 3.1 Summary of Experimental Dataset.

the number of videos	247
the number of genre labels	26
the number of key frames	9330
the vocabulary size of visual words	1000
the vocabulary size of speech words	6291

3.5 Experiments

In this section, the description of the experimental setup and data features will be introduced in detail. Then, I will discuss about the parameter tuning for the experimental system. Finally, the performances for the proposed model will be validated via both qualitative evaluation and quantitative evaluation.

3.5.1 Experiment Description

A video dataset provided by *MediaEval-2011 Genre Tagging Task*¹ [22] was used in this experiment. In their task, two types of data were extracted from videos. Firstly, a sequence of key frames deemed as image data were extracted from each video. Secondly, a number of speech transcript words deemed as speech data were tagged to each key frame for each video via an automatic speech recognition (ASR) method [22].

To convert image data into visual words, a number of individual SIFT descriptors [23, 24] were uniformly sampled from each key frame of each video, where the SIFT descriptor for every 10×10 pixel grid is computed conditioning in which the patch size was randomly sampled between scales of 10 to 30 pixels. Then, all the sampled SIFT descriptors from all the key frames of videos were clustered into k clusters through a k -means algorithm. These clustered individuals were treated as visual words with a vocabulary size of k . In this experiment, the vocabulary size of visual words was set to 1000. For the original speech transcript words, 418 types of standard stop words [25] and the speech words appearing in less than five videos were removed, and then the remained other 6291 types of common words (*i.e.*, the vocabulary size of speech words is 6291) were considered speech words that were used in the experiment.

¹Only *dev_data* dataset was utilized in this experiment. The source videos in the dataset were originally collected from the “blip.tv” video hosting service. See more details on homepage: <http://www.multimediaeval.org/mediaeval2011/>

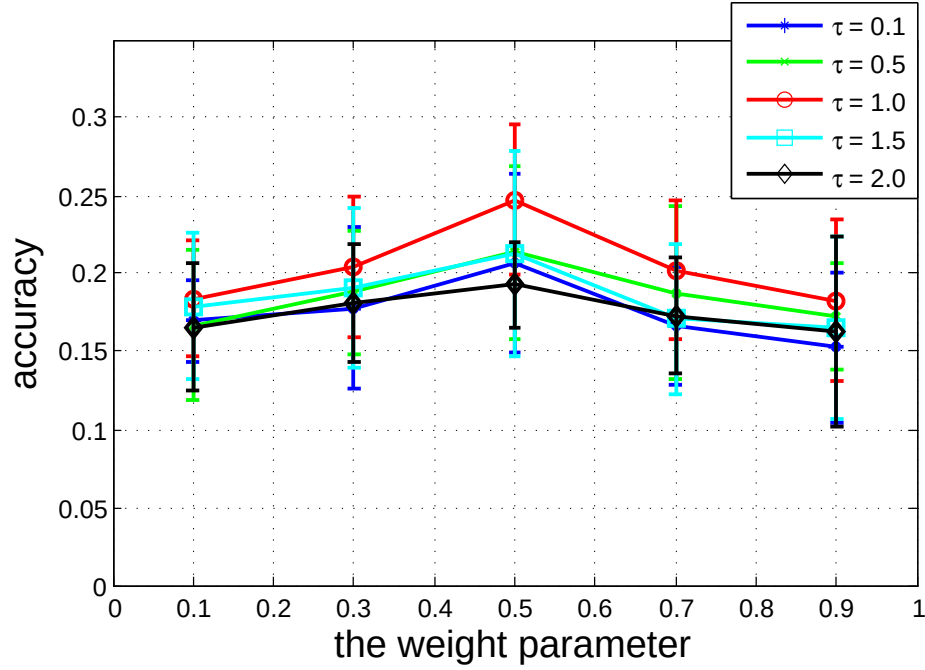


Fig. 3.5 Result for parameter tuning on Seq-cHDP with accuracy.

The summary of the experimental dataset is shown in Table 3.1. Note that each video within the dataset was originally tagged with a specific genre, there were total 26 genres specified in the experimental dataset. Hence, a genre classification task was performed to evaluate the proposed model and baseline model. In this process, a nested cross-validation mechanism was utilized in this experiment, where the video dataset were evenly divided into five subsets for the use of cross-validation. One subset was treated as the test set while the remaining four subsets were used for four-fold cross-validation. This procedure was totally repeated five times. During the testing phase, the system predicted the genre label for each video in the test set based on the training results by applying support vector machines (SVM)² [26]. In addition, accuracy (micro-F1) and macro-F1 as performance metrics were applied for evaluating the models.

3.5.2 Parameter Tuning

In Seq-cHDP, I define two controlling parameters w_j^f and τ , which play a very important role in modeling optimization. Here, a parameter tuning task was performed to optimize the

²I used LIBSVM available at <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>. The procedure of genre prediction is described as follows: (1) I estimate topic distributions by Seq-cHDP for both training and test video datasets. (2) I then learn SVM using the topic distributions for training video dataset and the corresponding genre labels. (3) I finally predict a genre label for each test video using the SVM trained at step (2).

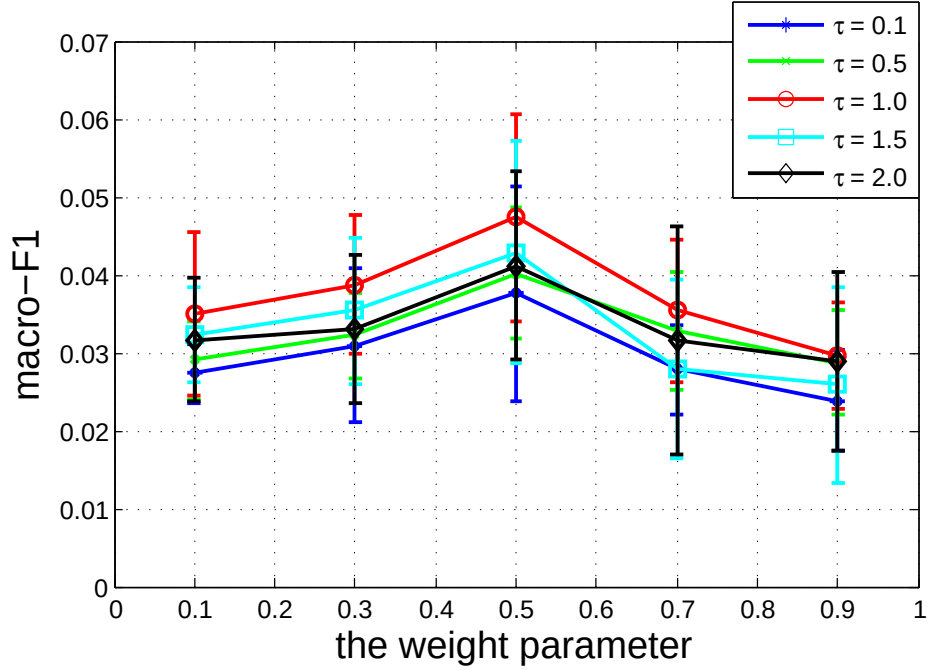


Fig. 3.6 Result for parameter tuning on Seq-cHDP with macro-F1.

Seq-cHDP before evaluating the performance of the Seq-cHDP. For simplicity, I consider all the weights w_j^f are equal to the same parameter w .

I respectively swept the parameter w and τ in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\{0.1, 0.5, 1.0, 1.5, 2.0\}$ during the experimental phase. The parameter tuning task was performed by using the validation set, which was separated from the training set. According to a *nested* cross-validation experiment, in each round of the four-fold cross validation, I can draw one of the four subsets as the validation set, which means I can conduct four different experiments by testing four individual validation sets separately. Hence, throughout all the five rounds in this cross-validation experiment, I can totally collect 5×4 sets of experimental results. For the initialization of Seq-cHDP, I set all the concentration parameters were drawn from a gamma prior $Ga(1.0, 1.0)$. The cascaded Gibbs sampling system stopped the iteration until all the variables within Seq-cHDP converged.

The results of parameter tuning is shown in Fig. 3.5 and Fig. 3.6, which demonstrate that the variation of the controlling parameters w and τ significantly affects the performance of Seq-cHDP with both of accuracy and macro-F1. The Seq-cHDP model achieved the best performance when initializing with $w = 0.5$ and $\tau = 1.0$.

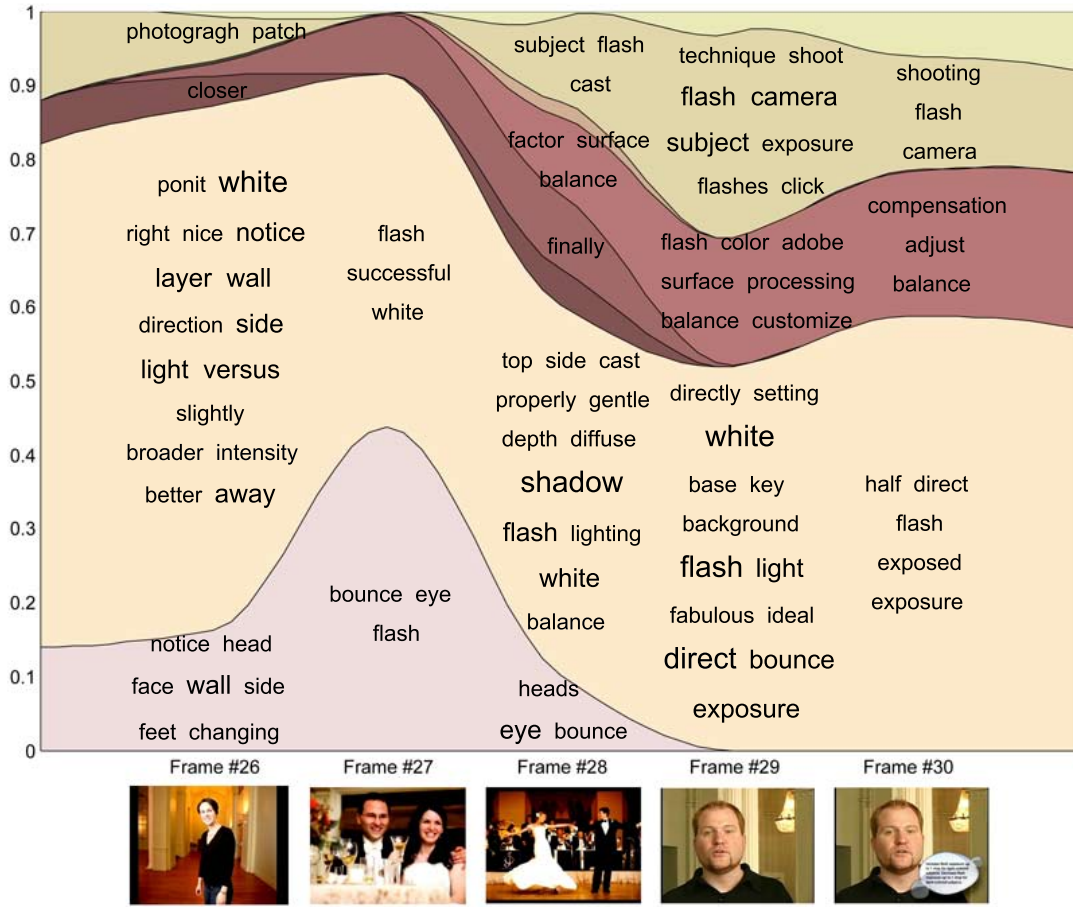


Fig. 3.7 Area graph for topic distribution in video No.109 (©Span-TV), superimposed with speech words.

3.5.3 Evaluation of Models

Both qualitative evaluation and quantitative evaluation on Seq-cHDP will be shown in this section. For comparison, three types of topic models were adopted as baseline models: the first one is a correspondence HDP (CorrHDP) model that is equivalent to the Seq-cHDP model conditioned on $w_j^f = 0$, the second one is a Seq-HDP model that is mentioned in Section 3.2, and the third one is correspondence LDA (CorrLDA) [14], whose number of topics must be predefined beforehand. For parameter settings, I referred to the results of the parameter tuning task shown in Section 3.5.2, and set $w = 0.5$ and $\tau = 1.0$.

Trend Estimation for Latent Topics

According to the nature of time dependent topic models, one of the advantages of Seq-cHDP is that it can estimate and track the trend of the latent topics along with time varying, while

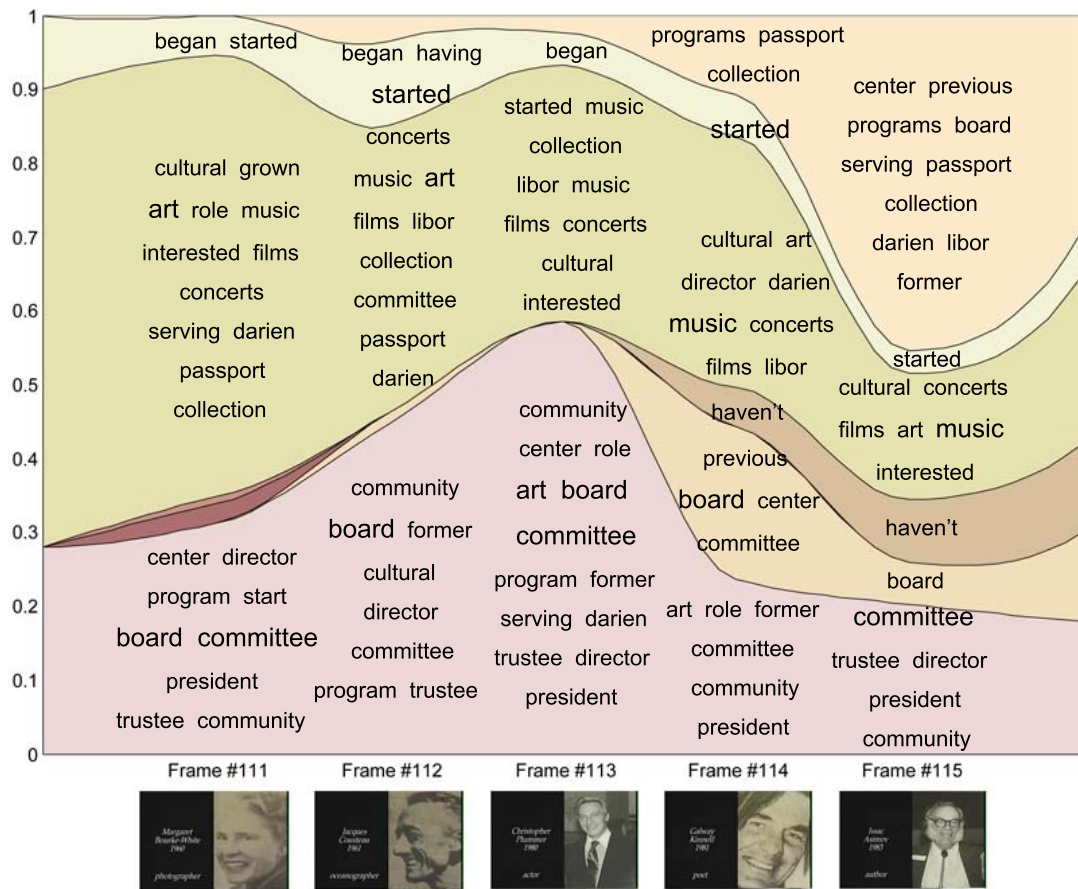


Fig. 3.8 Area graph for topic distribution in video No.221 (©Blue Orchid Productions), superimposed with speech words.

other non-time-dependent topic models fail to do so, since no time-dependency mechanism is incorporated in their models.

To demonstrate the performance of topic trend estimation with Seq-CHDP, a set of frame-based area graphs of topic distribution on three different examples that occurred in three corresponding videos are plotted in Fig. 3.9, Fig. 3.7 and Fig. 3.8, respectively. In this figure, each colored stripe represents a topic superimposed with assigned speech words, and the width of these stripes on the y-axis indicates the topic distribution over visual words. The x-axis represents the sequential key frames. Considering the space limitation, only some key speech words are shown on the area graph. In addition, the font size of these key speech words is proportional to their appearance frequencies in each key frame.

Based on the analysis to all these three examples, it is obvious to see that the topic distribution smoothly was varying along with key frame evolving because of the time-dependency mechanism in Seq-CHDP. Furthermore, through an adequate learning, the Seq-

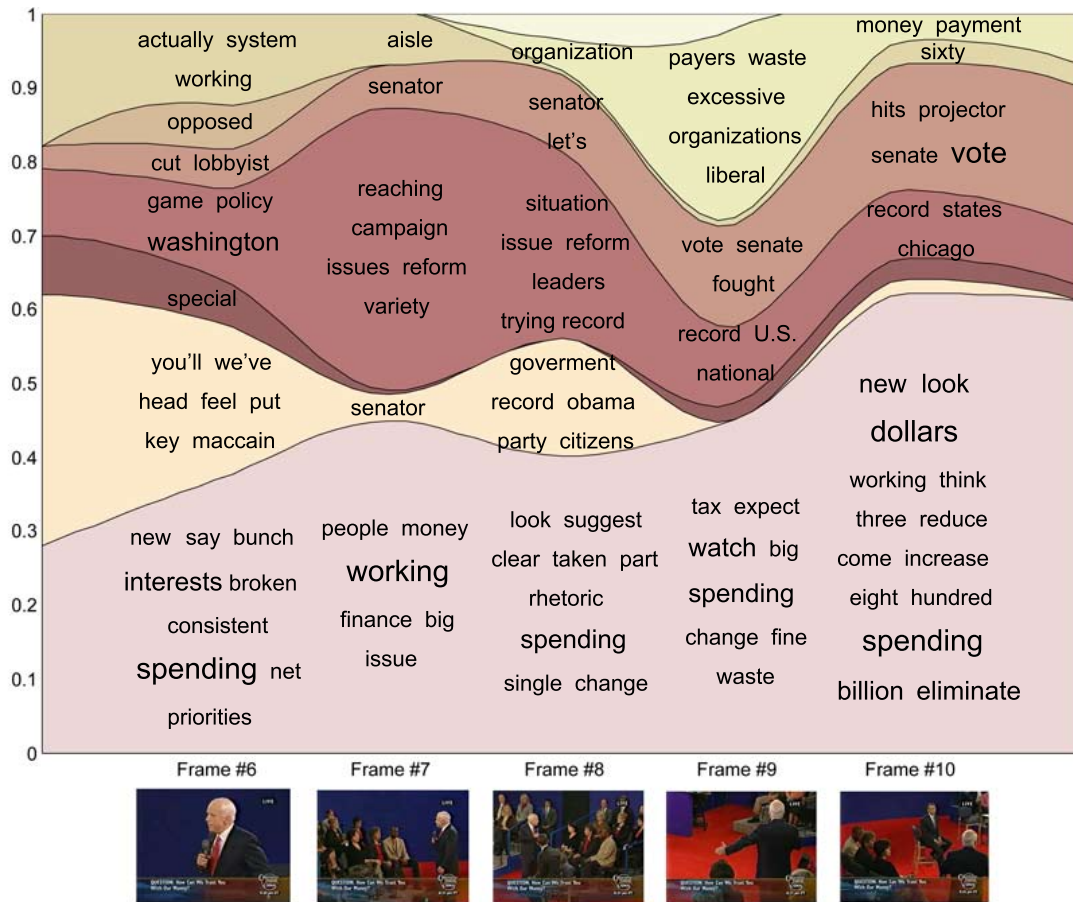


Fig. 3.9 Area graph for topic distribution in video No.106 (©Darien Library), superimposed with speech words.

cHDP model effectively grouped a number of speech words that have close meanings or belong to a similar class into a specific topic. For instance in Fig. 3.9, the video No.106 narrated a US presidential candidate debate on the question “Can we trust you with our money?”. In this case, I found that the major topic colored in pink collected a large number of semantically similar words such as *finance*, *dollars*, *spending*, *tax*, which implies that this topic is probably related to the key word “money”. Another example in Fig. 3.7, the video No.109 addressed four signature lighting techniques. As a consequence, the key speech words *flash*, *shadow*, *white*, *light* frequently showed in the major topic colored in yellow.

Through this task, we can have a clear insight into the evolving process of the latent topics within a video, and this is of benefit to the video semantic analysis.

Table 3.2 Performance of Genre Classification.

	Accuracy	Macro-F1
Seq-cHDP	0.2515 ± 0.0365	0.0558 ± 0.0122
CorrHDP	0.1903 ± 0.0219	0.0367 ± 0.0106
Seq-HDP(Visual)	0.2012 ± 0.0134	0.0388 ± 0.0066
Seq-HDP(speech)	0.1826 ± 0.0221	0.0349 ± 0.0121
CorrLDA(K=10)	0.1831 ± 0.0197	0.0358 ± 0.0059
CorrLDA(K=20)	0.1873 ± 0.0258	0.0379 ± 0.0112
CorrLDA(K=40)	0.1791 ± 0.0313	0.0355 ± 0.0128

Performance of Genre Classification

Here, I performed a genre classification task to quantitatively evaluate the target model. As mentioned in the beginning of this section, there were three baseline models, *i.e.*, CorrHDP, Seq-HDP and CorrLDA, which were utilized to compare with Seq-cHDP. For the initialization of baseline models, the controlling parameter τ in CorrHDP was also set to 1, while the controlling parameter w in Seq-HDP was also set to 0.5. The experiments performed with Seq-HDP had two different scenarios, which were respectively conducted by using single visual data and single speech data, since Seq-HDP is a unimodal topic model. For the CorrLDA model, I respectively conducted three sets of the experiments with different number of topics $K = 10$, $K = 20$ and $K = 40$. All the models were evaluated with a five-fold cross validation.

Table. 3.2 shows the performance of genre classification among all the topic models. As we can see, the accuracy and Macro-F1 results of Seq-cHDP achieved 0.2515 and 0.0558 respectively, which outperformed those of the other baselines shown in the table. In these results, we can see that the performance measured by accuracy is more significant than the performance measured by Macro-F1 for all the models, probably because the ground-truth genre distribution on experimental videos is imbalanced. Besides, CorrHDP performed at almost the same level as CorrLDA initialized with the most appropriate K , but still performed worse than Seq-cHDP, therefore I speculate that incorporating the time-dependency mechanism can make the model work more effectively. Seq-HDP no matter with single visual data or single speech data showed poorer performance than Seq-cHDP under the same condition, which demonstrated the effectiveness of the correspondence method for modeling multimodal data in Seq-cHDP.

3.6 Summary

In this chapter, I proposed a sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) model extended from HDP to analyze the semantic information of video data.

Seq-cHDP has a three-layer hierarchy in accordance with the structure of video data. In this work, two major contributions were made on Seq-cHDP: (1) a time-dependency mechanism based on a Markovian assumption was developed to model the time relativity between neighboring video frames, (2) a data correspondence mechanism was incorporated to handle the multimodal data issue within videos. To infer the Seq-cHDP, an inference method based on posterior representation sampler was derived to provide a way to estimate every variable within the model, and implemented via a cascaded Gibbs sampling scheme. The Seq-cHDP was experimentally evaluated by showing the trend estimation of latent topics with video examples and the performance of genre classification. The experimental results demonstrated that Seq-cHDP outperformed other baseline topic models in terms of both accuracy and macro-F1.

In conclusion, Seq-cHDP as a Bayesian nonparametric topic model can effectively handle the issue of video semantic analysis. However, I still notice some limitations existed in Seq-cHDP. For example, the correspondence mechanism incorporated in Seq-cHDP need to determine the pivot data type (modality) among multimodal data beforehand, which hurts the flexibility of the model. This problem will be discussed and improved in Chapter 4.

Chapter 4

Sequential Symmetric Correspondence Hierarchical Dirichlet Processes for Video Semantic Analysis

4.1 Overview

In Chapter 3, I introduced a sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) model to analyze the semantic information for video data. However, there are two drawbacks with respect to the correspondence mechanism incorporated in Seq-cHDP: (1) the pivot data type (modality) needs to be manually specified beforehand, (2) the pivot must be fixed throughout the whole generative process. These drawbacks might hurt the flexibility of modeling multimodal data when using the Seq-cHDP.

To tackle this problem, this chapter focuses on exploring a way to further improve the multimodal data modeling method for videos in Bayesian nonparametric topic models. Based on the framework of sequential HDP (Seq-HDP) proposed in Chapter 3, I propose a more flexible nonparametric topic model called sequential symmetric correspondence HDP (Seq-Sym-cHDP) with a symmetric correspondence mechanism, which can spontaneously choose the pivot data type with a multinomial pivot flag generator when generating a topic for a corresponding word within each modality. The dependencies among all the data types of multimodal data are effectively balanced via such pivot assignments within Seq-Sym-cHDP. I still derive the posterior representation samplers [20] to infer the Seq-Sym-cHDP model, and implement it with a cascaded Gibbs sampling method. The experimental results

comprehensively demonstrate the Seq-Sym-cHDP outperforms other baseline nonparametric topic models with both qualitative evaluation and quantitative evaluation.

The rest of this chapter is organized as follows. Section 4.2 gives a brief review of the preliminary work on Seq-HDP. Section 4.3 firstly proposes a sequential conditionally independent HDP (Seq-CI-HDP) model inspired by [27] to model the multimodal data in a conditionally independent manner, then reviews the unidirectional correspondence mechanism for multimodal data modeling in Seq-cHDP, finally presents a Seq-Sym-cHDP model to provide the most appropriate and flexible solution for multimodal data modeling. Section 4.4 derives the posterior representation samplers to infer the Seq-Sym-cHDP. Finally, experimental evaluations and summary are respectively presented in Section 4.5 and Section 4.6.

4.2 Preliminary Work on Sequential HDP

Before introducing the Bayesian nonparametric multimodal topic modeling methods, I will give a brief review of the unimodal Seq-HDP, which is designed for modeling the hierarchy of video data.

As shown in Fig. 3.1 of Chapter 3, Seq-HDP is a three-layer HDP model. I define three different probability measures G , G_0^f and G_j^f to construct generative processes corresponding to different layers in Seq-HDP. G located on the top layer of Seq-HDP stands for an overall measure among all videos. Next, G_0^f stands for the global measure reflecting the video f on the second layer of Seq-HDP. On the third layer, G_j^f stands for the local measure corresponding to the frame j in the video f . Since video frames in the same video are sequential and time dependent, a Markovian transfer link with a transfer weight w_j^f is thus established between two adjacent local measures G_{j-1}^f and G_j^f in Seq-HDP. The generative process of Seq-HDP is as follows:

$$\begin{aligned} G &\sim DP(\xi, H) \\ G_0^f &\sim DP(\gamma^f, G) \\ G_j^f &\sim DP(\alpha_0^f, w_{j-1}^f G_{j-1}^f + (1 - w_{j-1}^f) G_0^f) \end{aligned} \tag{4.1}$$

where $DP()$ indicates a Dirichlet process, H is a base measure, ξ , γ^f and α_0^f are concentration parameters corresponding to G , G_0^f and G_j^f . Finally, each data sample x_{fji} can be drawn from a conditional probability distribution $f(x_{fji} | \theta_{fji})$ that is conditioned on the factor θ_{fji} , and θ_{fji} is directly drawn from the local measure G_j^f .

To make the model more understandable and applicable, Seq-HDP can be reformulated with a stick-breaking construction as shown in Fig. 3.2 of Chapter 3, where three different component weight vectors $\mathbf{v}$, β^f , and π_j^f are utilized to construct corresponding measures in the different layers of Seq-HDP, the process is described as:

$$G = \sum_{k=1}^{\infty} v_k \delta_{\phi_k}, \quad G_0^f = \sum_{k=1}^{\infty} \beta_k^f \delta_{\phi_k}, \quad G_j^f = \sum_{k=1}^{\infty} \pi_{jk}^f \delta_{\phi_k} \quad (4.2)$$

where each component weight vector comprises an infinite number of corresponding component weights, *i.e.*, $\mathbf{v} = \{v_1, v_2, \dots, v_k, \dots\}$, $\beta^f = \{\beta_1^f, \beta_2^f, \dots, \beta_k^f, \dots\}$, and $\pi_j^f = \{\pi_{j1}^f, \pi_{j2}^f, \dots, \pi_{jk}^f, \dots\}$. Note that the components $\phi = \{\phi_k\}_{k=1}^{\infty}$ are distributed on the base measure H .

Following the hierarchy of Seq-HDP, I can successively draw these component weights with:

$$\mathbf{v} \sim GEM(\xi), \quad \beta^f \sim DP(\gamma^f, \mathbf{v}), \quad \pi_j^f \sim DP(\alpha_0^f, \pi_{j-1}^f) \quad (4.3)$$

where $GEM()$ represents a GEM distribution [19]. Note that the generation of π_j^f is influenced by both β^f and π_{j-1}^f because of the time dependency. Hence, π_j^f is expressed as follows.

$$\pi_j^f = w_{j-1}^f \pi_{j-1}^f + (1 - w_{j-1}^f) \beta^f \quad (4.4)$$

Then, for each data sample, the component index z_{fji} is directly drawn from π_{jk}^f . The data sample x_{fji} is generated from a conditional probability distribution $f(x_{fji} | \phi_{z_{fji}})$ conditioned on the sampled z_{fji} . In the case of a real application, z_{fji} can be deemed as a sampled latent topic, and x_{fji} is a generated word.

4.3 Multimodal Data Modeling Methods

A video generally includes image information and speech information from the perspective of video semantic analysis, which can be deemed as a type of multimodal data. This section focuses on how to deal with the issue of multimodal data modeling on the basis of Seq-HDP. Firstly, borrowing the ideas from conditionally independent LDA (CI-LDA) [27] and CorrLDA [14], I introduce two simple methods Seq-CI-HDP and Seq-cHDP to model the multimodal data, and discuss their features and drawbacks. Then, I propose a Seq-Sym-cHDP model to improve the issues mentioned in Seq-CI-HDP and Seq-cHDP via using a symmetric correspondence mechanism.

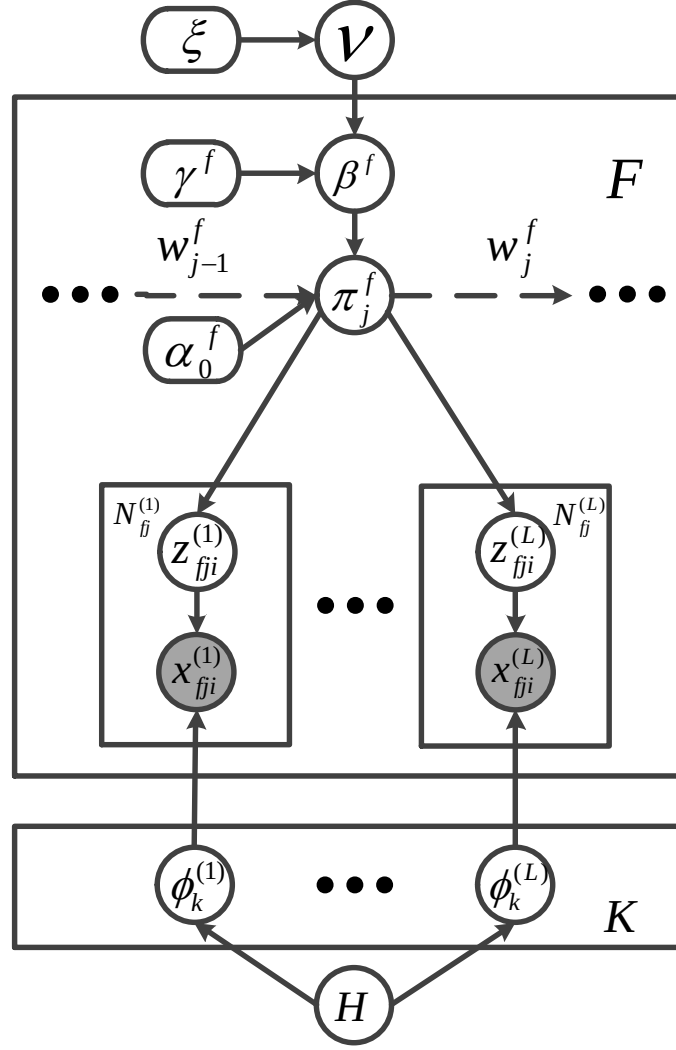


Fig. 4.1 Simplified Graphical model representation for stick-breaking construction of Seq-CI-HDP.

4.3.1 Conditionally Independent Modeling Method

According to the theory on CI-LDA [27], it assumes that all the types of data (modalities) are conditionally independent for generating their own data samples, however, only their assigned topics are drawn from the same multinomial distribution over topics for each document. Inspired by this idea, I incorporate a conditionally independent mechanism into the Seq-HDP model for handling multimodal data, and name this new model with sequential conditionally independent hierarchical Dirichlet processes (Seq-CI-HDP).

Fig. 4.1 depicts a simplified graphical model representation for the stick-breaking construction of Seq-CI-HDP. In this figure, I use l to denote the index of data type (modality),

and $l \in \{1, \dots, L\}$. $N_{fj}^{(l)}$ indicates the total counts for the words of data type l in frame j of video f . Since Seq-CI-HDP is a type of multimodal version of Seq-HDP, here I only describe the generative process for multimodal data in Seq-CI-HDP as below. The rest of the generative process can be referred to the case of Seq-HDP in Section 4.2.

- **Step 1:** For the word i of data type l in frame j of video f , draw a topic $z_{fji}^{(l)}$ from π_j^f .
- **Step 2:** Conditioned on sampled topic $z_{fji}^{(l)}$, a word $x_{fji}^{(l)}$ is drawn from $f(x_{fji}^{(l)} | \phi_k^{(l)}, k = z_{fji}^{(l)})$.

Note that $f(x_{fji}^{(l)} | \phi_k^{(l)})$ denotes a word distribution conditioned on $\phi_k^{(l)}$. In fact, all the $\{\phi_k^{(l)}\}_{l=1}^L$ drawn from base measure H share the same $K(\infty)$ topics. In addition, I notice that the sampling processes for all the topics are only related with π_j^f , which implies that the dependencies among all the multimodal data are relatively weak.

4.3.2 Correspondence Modeling Method

Review to Chapter 3, I proposed a Seq-cHDP model to handle the multimodal data in videos with a data correspondence mechanism. As opposed to Seq-CI-HDP, the data correspondence mechanism incorporated in Seq-cHDP specifies two different types of data: pivot data and non-pivot data, and establish a unidirectional dependency that makes non-pivot data draw their latent topics from the topic distribution over pivot data.

Similarly, I also show a simplified graphical model representation for the stick-breaking construction of Seq-cHDP in Fig. 4.2, and describe its generative process for multimodal data as below.

- **Step 1:** Assume that the data type $\{l\}_{l=1}$ is the pivot data type. Then a topic $z_{fji}^{(1)}$ for the word $x_{fji}^{(1)}$ is drawn from π_j^f . Conditioned on sampled $z_{fji}^{(1)}$, a specific word $x_{fji}^{(1)}$ is drawn from $f(x_{fji}^{(1)} | \phi_k^{(1)}, k = z_{fji}^{(1)})$.
- **Step 2:** For the other non-pivot data types $\{l\}_{l=2}^L$, a topic $z_{fji}^{(l)}$ for the word $x_{fji}^{(l)}$ is drawn from $Uniform(z_{fj1}^{(1)}, \dots, z_{fjN_{fj}^{(1)}}^{(1)})$. Conditioned on sampled $z_{fji}^{(l)}$, a word $x_{fji}^{(l)}$ is drawn from $f(x_{fji}^{(l)} | \phi_k^{(l)}, k = z_{fji}^{(l)})$.

Here, $Uniform()$ represents a uniform distribution, for the explanation of other parameters, we can refer to the similar parameters described in Seq-CI-HDP. As discussed in Section 3.3 of Chapter 3, the topic assignment for non-pivot data completely depends on the topic distribution over predefined pivot data. However, such unidirectional dependency for modeling multimodal data makes the model lack flexibility, for instance, the model might show poor performance when the data in some non-pivot data types become more dominant than the data in the pivot data type.

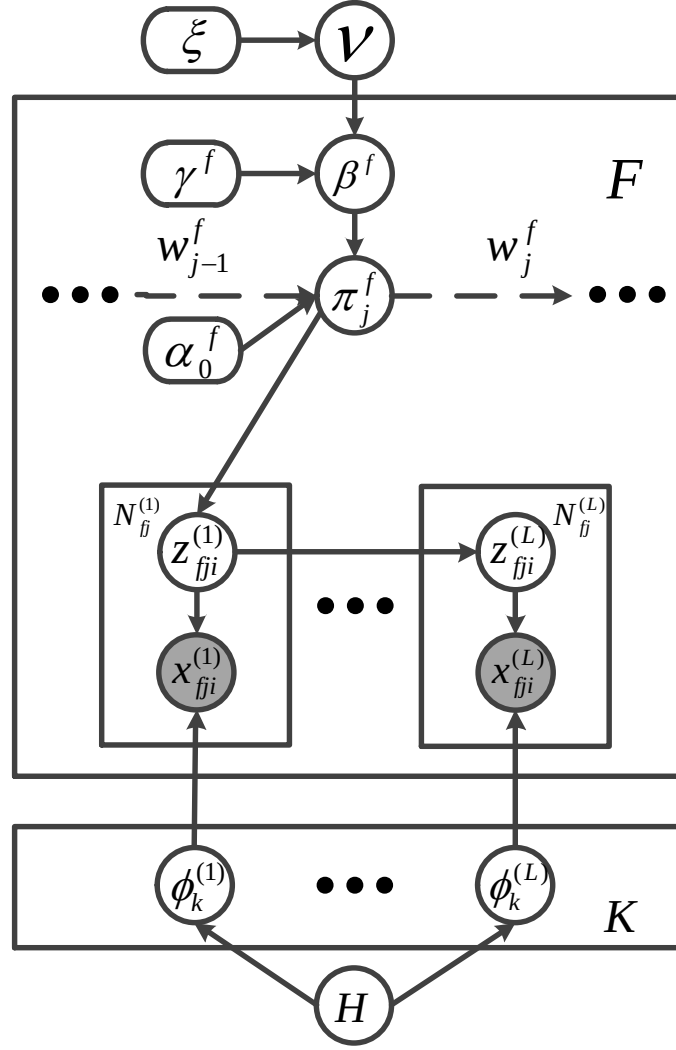


Fig. 4.2 Simplified Graphical model representation for stick-breaking construction of Seq-cHDP.

4.3.3 Symmetric Correspondence Modeling Method

In Section 4.3.1 and Section 4.3.2, Seq-CI-HDP and Seq-cHDP are respectively introduced for modeling multimodal data. However, these two methods have some drawbacks in terms of their mechanisms. More specifically, Seq-CI-HDP handles the generative process for each data type independently conditioned on latent topics, but the dependencies among all the multimodal data are relatively weak. Seq-cHDP can take care of the dependency issue for multimodal data via predefining a fixed pivot data type, however, such correspondence mechanism lacks flexibility.

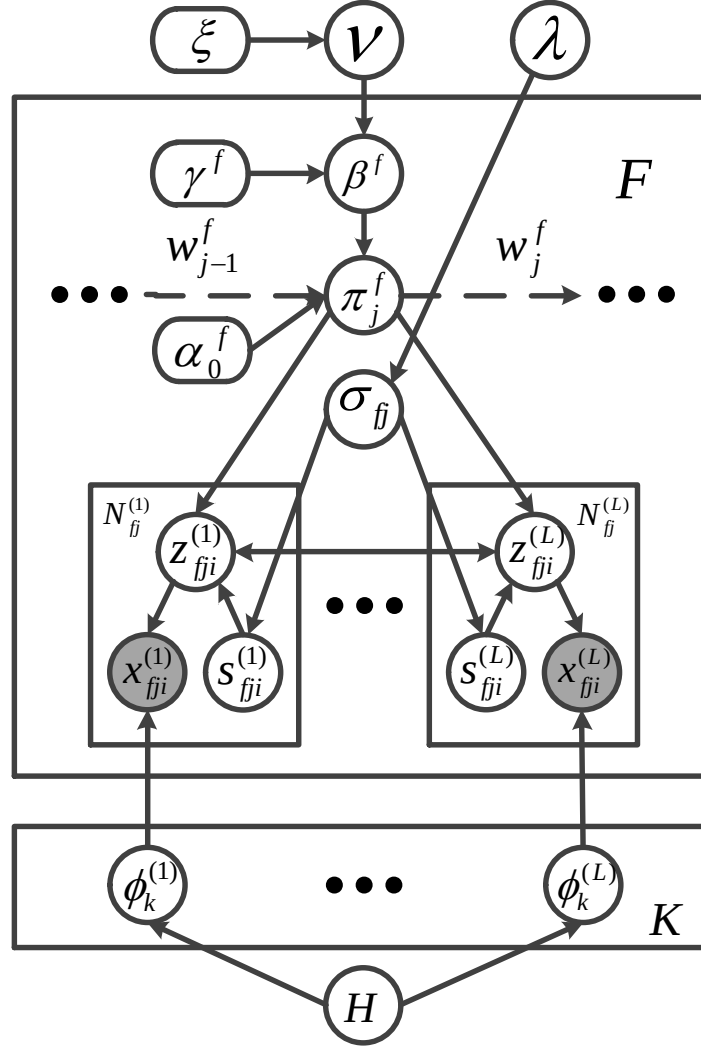


Fig. 4.3 Simplified Graphical model representation for stick-breaking construction of Seq-Sym-cHDP.

Consequently, I propose a more flexible model called sequential symmetric correspondence hierarchical Dirichlet processes (Seq-Sym-cHDP) to improve the multimodal data modeling method. Seq-Sym-cHDP incorporates a symmetric correspondence mechanism that can spontaneously select the pivot data type when generating a topic for a corresponding word. A similar idea also appeared in previous study [15], where the symmetric correspondence method is proposed on the basis of LDA for multilingual text analysis. Fig. 4.3 depicts a simplified graphical model representation for the stick-breaking construction of Seq-Sym-cHDP. The generative process for multimodal data within Seq-Sym-cHDP is as below.

- **Step 1:** Draw a multinomial pivot flag generator σ_{fj} from $Dir(\lambda)$.
- **Step 2:** For the word i of data type l in frame j of video f , draw a pivot flag $s_{fji}^{(l)}$ from $Multinomial(\sigma_{fj})$.
- **Step 3:** If $s_{fji}^{(l)} = l$, draw a topic $z_{fji}^{(l)}$ from π_j^f . Otherwise if $s_{fji}^{(l)} = h \neq l$, draw a topic $z_{fji}^{(l)}$ from $Uniform(z_{fj1}^{(h)}, \dots, z_{fjH_{fj}^{(h)}}^{(h)})$.
- **Step 4:** Conditioned on sampled $z_{fji}^{(l)}$, a word $x_{fji}^{(l)}$ is drawn from $f(x_{fji}^{(l)} | \phi_k^{(l)}, k = z_{fji}^{(l)})$.

Here, $Dir(\lambda)$ indicates a Dirichlet distribution with a hyperparameter λ , and $Multinomial()$ indicates a multinomial distribution. For the other parameters, we can get explanations from the description in Seq-CI-HDP. As shown in this process, the multinomial pivot flag generator σ_{fj} generates a specific pivot flag to choose the current pivot data type every time before sampling a topic for the current word, and the generated pivot data type affects the way of sampling the topic. For the word $x_{fji}^{(l)}$ within data type l , if its sampled pivot flag points at the same data type (i.e., $s_{fji}^{(l)} = l$), the process of its topic assignment will be independent from other topic assignments and only associated with π_j^f . However, if its sampled pivot flag points at the other data type h (i.e., $s_{fji}^{(l)} = h \neq l$), its topic will be drawn from a uniform distribution that consists of all the topics assigned to $H_{fj}^{(h)}$ words within data type h . Distinguished from $N_{fj}^{(h)}$, $H_{fj}^{(h)}$ indicates the number of words that have been already assigned with topics at the current step. Hence, there is $H_{fj}^{(h)} \leq N_{fj}^{(h)}$.

Based on such a symmetric correspondence method, Seq-Sym-cHDP can flexibly balance the dependencies among all the multimodal data via managing the pivot flag assignment. The symmetric correspondence method can still make the model work effectively, even if when facing more complex scenario of multimodal data. For example in the case of video data modeling, when the number of the speech words (or visual words) drastically decreases and the number of the visual words (or speech words) remains almost unchanged in some frames of the video, the speech words (or visual words) might fail to express the entire semantical information for the current frames of this video, since the number of the speech words (or visual words) is insufficient. In this scenario, the symmetric correspondence method can still make the Seq-Sym-cHDP work well via balancing the dependencies between the speech words and the visual words with pivot assignments. In other words, Seq-Sym-cHDP can let visual words (or speech words) enriched with entire information be informatively complementary to these insufficient speech words (or visual words) via assigning pivot flags to some of the visual words (or speech words).

4.4 Inference Method for Seq-Sym-cHDP

In this section, I continue to use posterior representation sampler to infer the Seq-Sym-cHDP, and implement it with a cascaded Gibbs sampling method.

4.4.1 Posterior Representation Sampler

Since the Seq-Sym-cHDP model has a similar hierarchy with Seq-cHDP, the sampling scheme for component weights $\mathbf{v}$, β^f and π_j^f have a similar form with the one in Seq-cHDP, as described in Section 3.4. Therefore, $\mathbf{v}$, β^f and π_j^f can be respectively estimated as follows:

$$\begin{aligned}\mathbf{v} &= (v_1, \dots, v_K, v_u) \sim \text{Dir}(M_1, \dots, M_K, \xi) \\ \beta^f &= (\beta_1^f, \dots, \beta_K^f, \beta_u^f) \sim \text{Dir}(\tilde{\beta}_1^f, \dots, \tilde{\beta}_K^f, \tilde{\beta}_u^f) \\ \pi_j^f &= (\pi_{j1}^f, \dots, \pi_{jK}^f, \pi_{ju}^f) \sim \text{Dir}(\tilde{\pi}_{j1}^f, \dots, \tilde{\pi}_{jK}^f, \tilde{\pi}_{ju}^f)\end{aligned}\quad (4.5)$$

where all the component weight vectors are reformulated with a new augmentable finite vector that includes K components and a promising component u . For the estimator $\tilde{\beta}_k^f$ and $\tilde{\beta}_u^f$, they are computed with $\tilde{\beta}_k^f = \gamma^f v_k + T_{fk}^{0 \rightarrow \cdot}$ and $\tilde{\beta}_u^f = \gamma^f v_u$. However, for the estimator $\tilde{\pi}_{jk}^f$ and $\tilde{\pi}_{ju}^f$, they are computed with:

$$\begin{aligned}\tilde{\pi}_{jk}^f &= \alpha_0^f w_{j-1}^f \pi_{(j-1)k}^f + \alpha_0^f (1 - w_{j-1}^f) \beta_k^f + C_{fjlk}^{(\cdot)} + T_{fk}^{j \rightarrow j+1} \\ \tilde{\pi}_{ju}^f &= \alpha_0^f w_{j-1}^f \pi_{(j-1)u}^f + \alpha_0^f (1 - w_{j-1}^f) \beta_u^f\end{aligned}\quad (4.6)$$

where $C_{fjlk}^{(l)}$ is the count for the word with data type l in frame j of video f possessing the pivot flag with the same data type l conditioned on topic k , and $C_{fjlk}^{(\cdot)} = \sum_{(l)} C_{fjlk}^{(l)}$. Different from the computation of $\tilde{\pi}_{jk}^f$ defined in Seq-cHDP, not all the words are assigned topics directly from π_j^f in Seq-Sym-cHDP, we thus only need to count the number of the words whose assigned topics are straightforwardly drawn from π_j^f , i.e., the value of $C_{fjlk}^{(\cdot)}$.

For sampling the table counts T_{jk}^f and the metatable counts M_k^f of CRF, I use the same Bernoulli trial method [21], which is specified in Section 2.3 of Chapter 2. Note that the number of the customers is $C_{fjlk}^{(\cdot)}$ when sampling the T_{jk}^f in Seq-Sym-cHDP. Since $T_{jk}^f = T_{fk}^{j-1 \rightarrow j} + T_{fk}^{0 \rightarrow j}$, both $T_{fk}^{j-1 \rightarrow j}$ and $T_{fk}^{0 \rightarrow j}$ are sampled with a simple multinomial distribution, as described in Eq. 3.21 and Eq. 3.22 of Section 3.4.2 of Chapter 3.

The biggest difference on posterior representation sampler between Seq-cHDP and Seq-Sym-cHDP is topic sampling. In Seq-Sym-cHDP, the topic sampling is associated with the

pivot selection. I thus derive a full conditional joint likelihood for estimating both topic $z_{fji}^{(l)}$ and pivot flag $s_{fji}^{(l)}$ for each word as below:

$$\begin{aligned}
 P(z_{fji}^{(l)} = k, s_{fji}^{(l)} = l | x_{fji}^{(l)}, \lambda, \dots) &\propto P(s_{fji}^{(l)} = l | \lambda) P(z_{fji}^{(l)} = k | \pi_j^f) P(x_{fji}^{(l)} | z_{fji}^{(l)} = k, \dots) \\
 &= \begin{cases} \frac{C_{fjl}^{-fjli} + \lambda}{C_{fjl}^{-fjli} + \sum_{l' \neq l} C_{fjl'} + L\lambda} \cdot \pi_{jk}^f \cdot f_k^{-x_{fji}^{(l)}}(x_{fji}^{(l)}) & \text{if } k \text{ is used} \\ \frac{C_{fjl}^{-fjli} + \lambda}{C_{fjl}^{-fjli} + \sum_{l' \neq l} C_{fjl'} + L\lambda} \cdot \pi_{ju}^f \cdot f_{k^{new}}^{-x_{fji}^{(l)}}(x_{fji}^{(l)}) & \text{if } k \text{ is newborn} \end{cases} \quad (4.7)
 \end{aligned}$$

$$\begin{aligned}
 P(z_{fji}^{(l)} = k, s_{fji}^{(l)} = h | x_{fji}^{(l)}, \lambda, \dots) &\propto P(s_{fji}^{(l)} = h | \lambda) P(z_{fji}^{(l)} = k | Z_{fj}^{(h)}) P(x_{fji}^{(l)} | z_{fji}^{(l)} = k, \dots) \\
 &= \frac{C_{fjh}^{-fjli} + \lambda}{C_{fjh}^{-fjli} + \sum_{h' \neq h} C_{fjh'} + L\lambda} \cdot \frac{n_{fjk}^{(h)}}{N_{fj}^{(h)}} \cdot f_k^{-x_{fji}^{(l)}}(x_{fji}^{(l)}) \quad (4.8)
 \end{aligned}$$

where C_{fjl} indicates the count for the pivot flag pointing to data type l over all the multimodal data in frame j of video f , the superscript $-fjli$ for a variable denotes the terms excluding the $fjli$ th term. $Z_{fj}^{(h)}$ is a vector that records all the topic assignments for the words of data type h in frame j of video f , $n_{fjk}^{(h)}$ indicates the count for topic k assigned to words of data type h in frame j of video f , and $N_{fj}^{(h)}$ indicates the total count for the words in the same domain. Note that there are two kinds of conditional word likelihood functions ($f_k^{-x_{fji}^{(l)}}(x_{fji}^{(l)})$ and $f_{k^{new}}^{-x_{fji}^{(l)}}(x_{fji}^{(l)})$) conditioned on topic k . When the topic k is a previously used topic, $f_k^{-x_{fji}^{(l)}}(x_{fji}^{(l)})$ is formulated with:

$$\begin{aligned}
 f_k^{-x_{fji}^{(l)}}(x_{fji}^{(l)} = v^{(l)}) &= \int f(x_{fji}^{(l)} = v^{(l)} | \phi_k^{(l)}) p(\phi_k^{(l)} | X_{lk}^{-fjli}, H) d\phi_k^{(l)} \\
 &= \frac{n_{kv^{(l)}}^{-fjli} + \tau}{\sum_{v'^{(l)}} n_{kv'^{(l)}}^{-fjli} + V^{(l)} \tau} \quad (4.9)
 \end{aligned}$$

when the topic k is a newborn topic, $f_{k^{new}}^{-x_{fji}^{(l)}}(x_{fji}^{(l)})$ is formulated with:

$$f_{k^{new}}^{-x_{fji}^{(l)}}(x_{fji}^{(l)}) = \int f(x_{fji}^{(l)} = v^{(l)} | \phi_k^{(l)}) p(\phi_k^{(l)} | H) d\phi_k^{(l)} = \frac{1}{V^{(l)}} \quad (4.10)$$

where $v^{(l)}$ is the index of words in terms of the vocabulary in data type l . X_{lk}^{-fjli} is a vector that records all the words of data type l with topic k except for $x_{fji}^{(l)}$. $n_{kv^{(l)}}^{-fjli}$ is the counts for

Algorithm 3: Cascaded Gibbs Sampling for Inferring Seq-Sym-cHDP**Initialization;**

```

for ( $f = 1; f \leq F; f++$ ) do
  for ( $j = J^f; j \geq 1; j--$ ) do
    for ( $l = 1; l \leq L; l++$ ) do
      | Sample pivot flags  $S_{fj}^{(l)}$  and topics  $Z_{fj}^{(l)}$ ;
    end
    for ( $k = 1; k \leq K; k++$ ) do
      | Sample table counts  $T_{jk}^f, T_{fk}^{j-1 \rightarrow j}$  and  $T_{fk}^{0 \rightarrow j}$ ;
    end
  end
  for ( $k = 1; k \leq K; k++$ ) do
    | Sample metatable counts  $M_k^f$ ;
  end
end
Sample concentration parameters  $\xi, \gamma^f$  and  $\alpha_0^f$ ;
Sample overall component weights  $v$ ;
for ( $f = 1; f \leq F; f++$ ) do
  | Sample global component weights  $\beta^f$ ;
  for ( $j = 1; j \leq J^f; j++$ ) do
    | Sample local component weights  $\pi_j^f$ ;
  end
end

```

the word $v^{(l)}$ assigned to topic k except for $x_{fji}^{(l)}$. $V^{(l)}$ is the vocabulary size of data type l . τ is a hyperparameter for a Dirichlet distribution, and $H = \text{Dir}(\tau)$.

The sampling methods of hyperparameters ξ, γ^f and α_0^f are the same with the ones in Seq-cHDP, which are also specified in Section 3.4.2. In addition, the hyperparameter λ is considered as another controlling parameter that will be determined in the parameter tuning stage. For simplicity, the hyperparameter τ is set to a constant value in this chapter.

4.4.2 Cascaded Gibbs Sampling

Based on the posterior representation sampler, I also devise a cascaded Gibbs sampling scheme to implement this inference method for Seq-Sym-cHDP, whose pseudo-code is shown in Algorithm 3.

Table 4.1 Summary of Experimental Dataset

the number of video files	1726
the number of genre labels	26
the number of key frames	40588
the number of visual words	166803k
the number of speech words	1318k
the vocabulary size of visual words	1000
the vocabulary size of speech words	17596

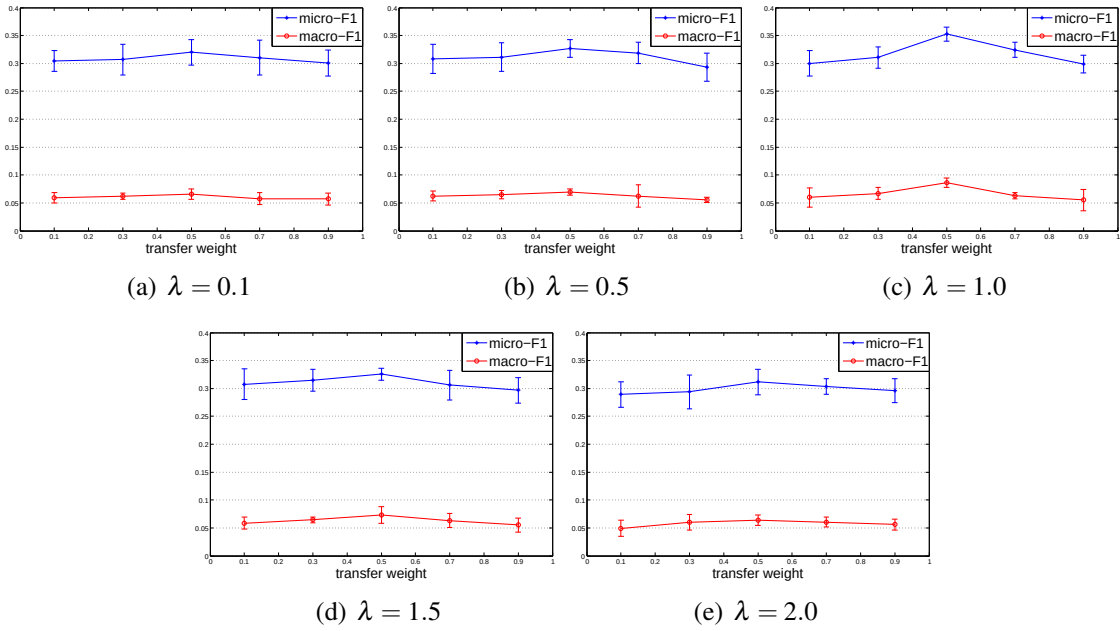


Fig. 4.4 Parameter tuning results for Seq-Sym-cHDP.

4.5 Experiments

4.5.1 Experimental Setup

In this section, I utilized a *test_data* dataset extracted from *MediaEval-2011 Genre Tagging Task* [22] to conduct this experiment. Compared with the *dev_data* dataset used in Chapter 3, the *test_data* collects more videos in its dataset, furthermore, the genre distribution on these videos within the *test_data* dataset is more balanced. Similarly, two types of data, *i.e.*, image data and speech data, were extracted from videos. For each video, the image data comprised a sequence of key frames, which were converted into visual words by computing and clustering SIFT descriptors [23, 24], while the speech data comprised a number of speech transcript

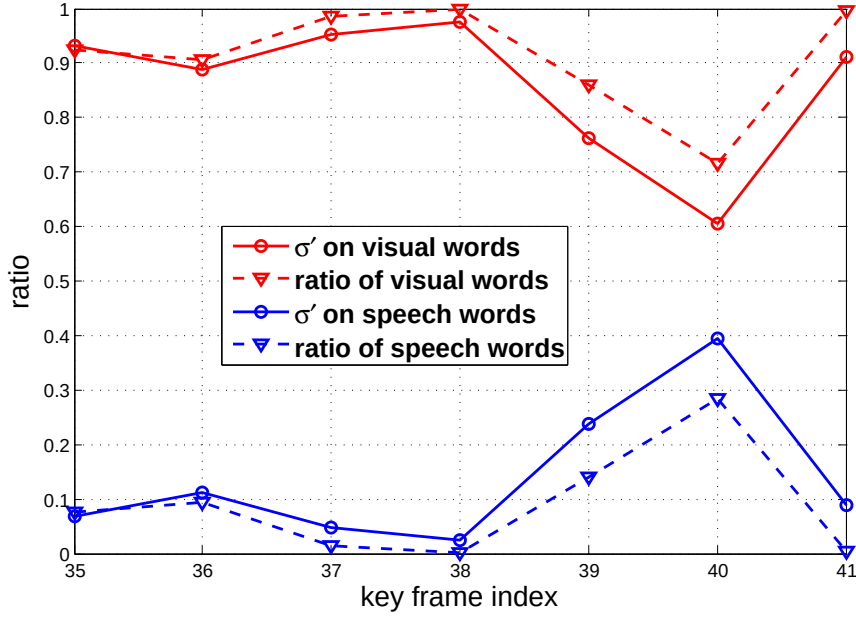


Fig. 4.5 Performance of pivot assignments in video No.161 with Seq-Sym-cHDP.

words, which were tagged to each key frame and converted into speech words by removing standard stop words [25] and rare words. This video preprocessing is identical to the one described in Section 3.5 of Chapter 3. The summary of the experimental dataset is shown in Table 4.1.

To comprehensively assess the model, both qualitative evaluation and quantitative evaluation are given in the experimental results. For quantitative evaluation, I still use the genre classification with a five-fold nested cross-validation scheme, as what was done in Section 3.5 of Chapter 3. Additionally, accuracy (micro-F1) and macro-F1 are utilized as performance metrics for quantitative evaluation.

4.5.2 Parameter Tuning

In Seq-Sym-cHDP, I predefine w_j^f and λ as controlling parameters, whose values need to be determined to optimize the model via a parameter tuning task. For simplicity, all the transfer weights w_j^f are assumed to be the same parameter w . To initialize the model, I specified that all the hyperparameters ξ , γ^f , and α_0^f were drawn from a gamma prior $\text{Gamma}(1.0, 1.0)$, and set τ to 1.0. I let the cascaded Gibbs sampling system run 1000 iterations in total so as to guarantee that all the variables in the model get fully converged.

During the experiment, I respectively swept the controlling parameter w and λ in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\{0.1, 0.5, 1.0, 1.5, 2.0\}$ and conducted this parameter tuning task

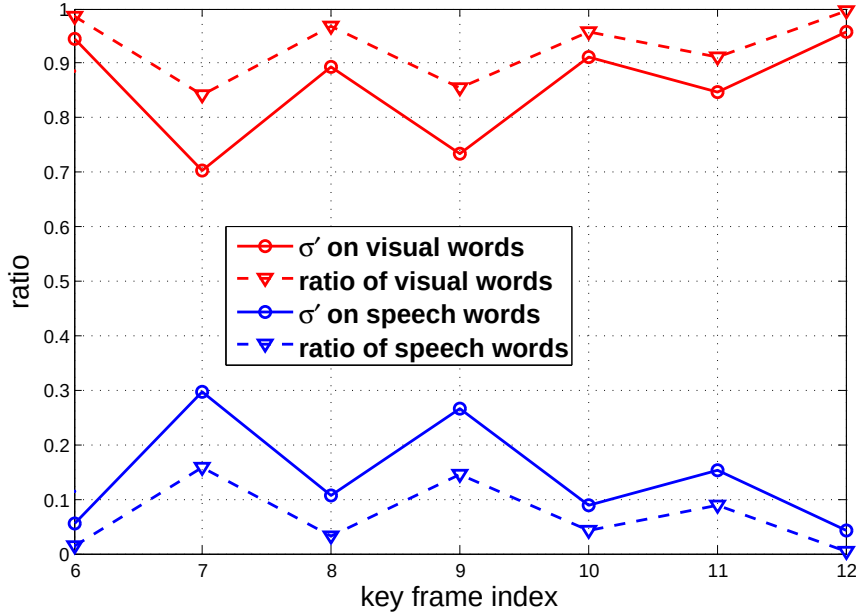


Fig. 4.6 Performance of pivot assignments in video No.511 with Seq-Sym-cHDP.

by using the validation set from training set in five-fold nested cross validation. Then, $25 \times 4 \times 5$ sets of experimental results were collected in total. The average parameter tuning results are shown in Fig. 4.4. According to both accuracy (micro-F1) results and macro-F1 results, the Seq-Sym-cHDP achieved the best performance with $w = 0.5$ and $\lambda = 1.0$.

4.5.3 Qualitative Evaluation

In this experiment, qualitative evaluation is introduced from two aspects: (1) dynamic pivot assignments, (2) trend estimation for latent topics. The parameter settings is properly set based on the results of parameter tuning, *i.e.*, $w = 0.5$ and $\lambda = 1.0$.

Dynamic Pivot Assignments

As mentioned in Section 4.3.3, the major advantage of Seq-Sym-cHDP is that it can flexibly balance the dependencies among all the multimodal data via managing the pivot flag assignment. I illustrate the performance of dynamic pivot assignments with three different examples occurring in three corresponding videos, as shown in Fig. 4.5, Fig. 4.6 and Fig. 4.7, respectively. Here, σ' indicates the the proportion of pivot assignments, for each data type l

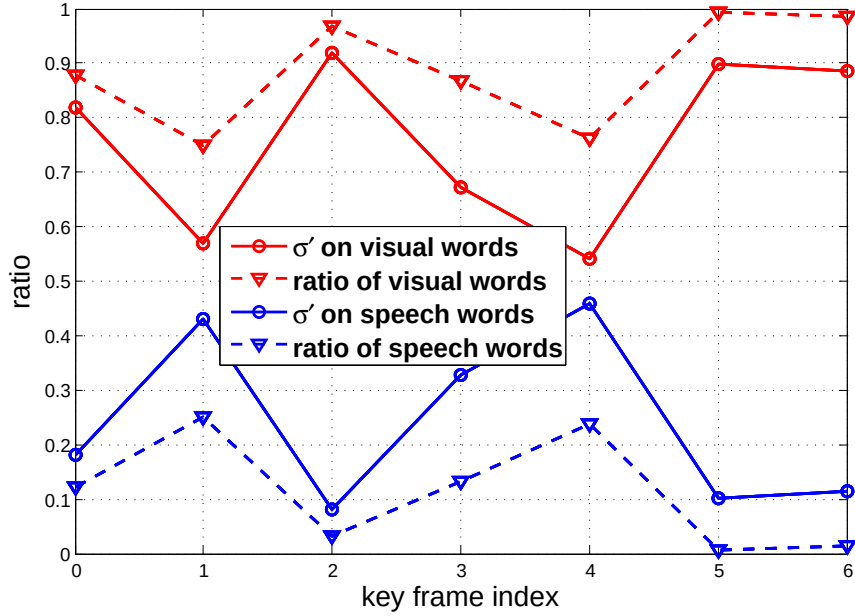


Fig. 4.7 Performance of pivot assignments in video No.984 with Seq-Sym-cHDP.

in key frame j of video f , the σ'_{fjl} is expressed as below.

$$\sigma'_{fjl} = \frac{C_{fjl}}{\sum_{l'} C_{fjl'}} \quad (4.11)$$

In the results, both the variation curves of σ' on visual words and speech words was plotted against key frames with solid lines, for comparison, the real ratios of the number of visual words and speech words among all the multimodal words was also plotted with dotted lines.

According to these examples, the count ratio of words for each data type was dynamically changing along with the variation of key frames. Mostly, the variation trend on σ' followed the variation trend on the ratio of the words within the same data type. In other words, the pivot assignments to the current data type were increased (or decreased), when the count ratio of the words within the current data type rose (or fell). These results demonstrate that the symmetric correspondence mechanism can make Seq-Sym-cHDP work more effectively and flexibly on balancing the dependencies among all the multimodal data.

Trend Estimation for Latent Topics

As mentioned in Section 3.5.3 of Chapter 3, predicting the variation trend of latent topics within a video is another important strength for all the time dependent topic models. Especially for Seq-Sym-cHDP, its particular symmetric correspondence mechanism guarantees

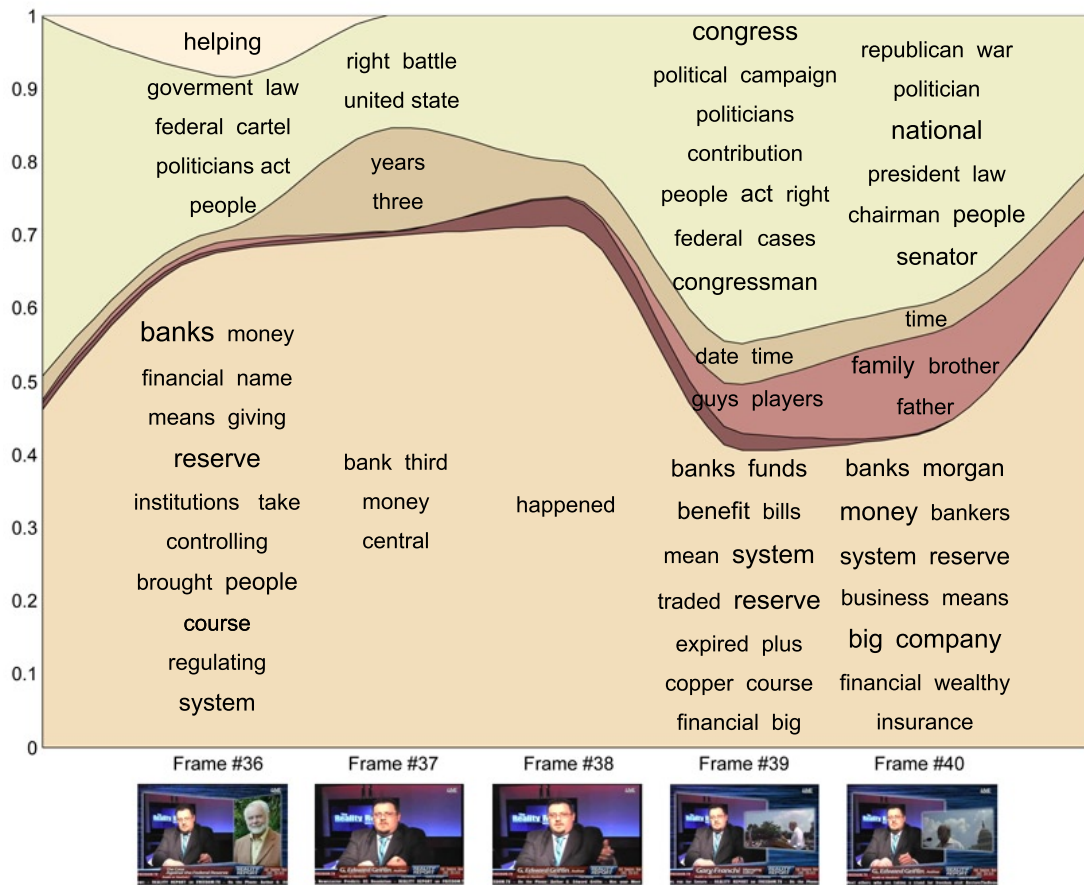


Fig. 4.8 Area graph for topic distribution in video No.161 (©The Reality Report) with Seq-Sym-cHDP, superimposed with speech words.

that the model can learn and track the variation trend of latent topics with more accuracy and stability.

Here, I still use the same three example video clips, and draw their area graphs of topic distributions to illustrate the performance of topic trend estimation with Seq-Sym-cHDP, as shown in Fig. 4.8, Fig. 4.9 and Fig. 4.10, respectively. where the topic trend within each video was effectively estimated with sequential topic distributions because of the time-dependency mechanism. For instance in video No.511, this clip is a talk show that focuses on a discussion about political issues, therefore the Seq-Sym-cHDP predicted that the major topic with key speech words such as *military*, *political*, *government*, *country* ran through all the key frames. Another example in video No.984, this clip is a tutorial on how to use photoshop to process an image, therefore the key speech words such as *move*, *shift*, *click*, *draw* that describe the operation with photoshop were clustered into the major topic by the Seq-Sym-cHDP.

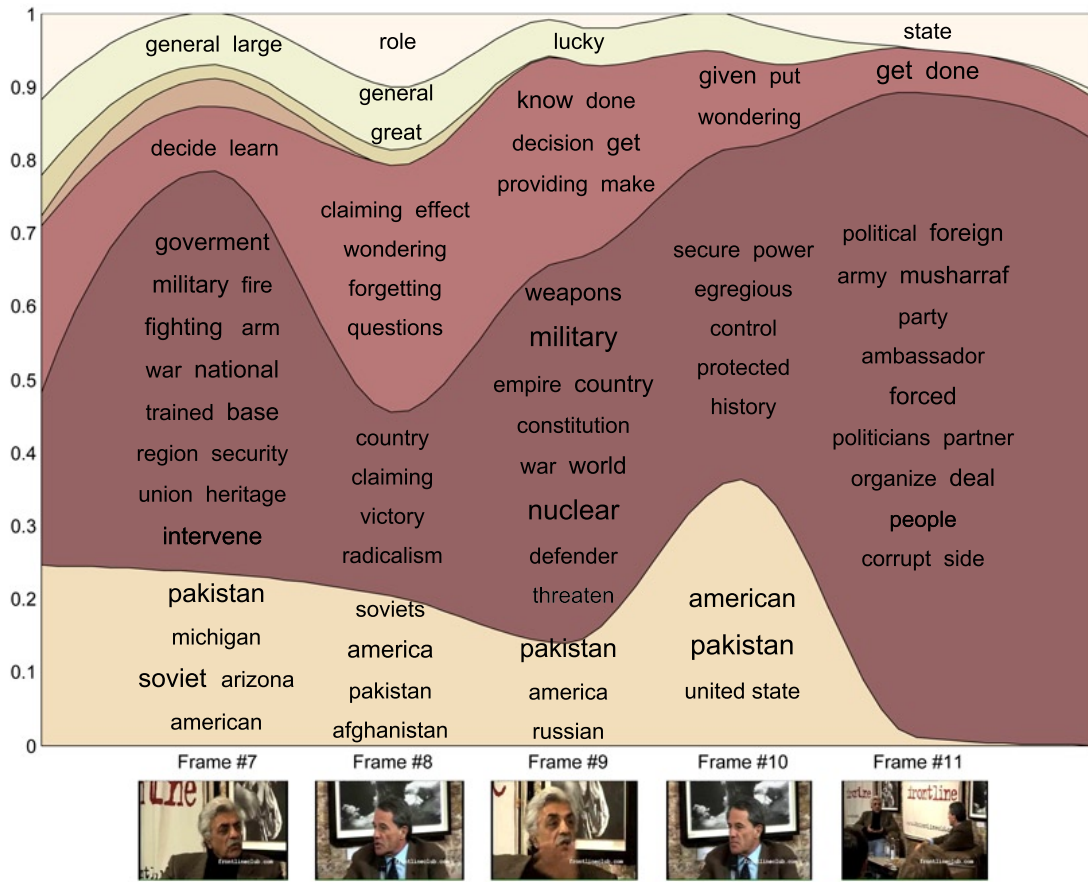


Fig. 4.9 Area graph for topic distribution in video No.511 (©Frontline Club) with Seq-Sym-cHDP, superimposed with speech words.

However, when encountering more extreme scenarios such as extremely imbalanced multimodal data, Seq-Sym-cHDP still can handle such tricky task well. In video No.161, the visual words were very dominant in Frame #37 and #38, since the number of speech words was very few. Similarly in video No.984, the same phenomenon occurred in Frame #2 and #5. In these scenarios, the speech words failed to express the entire semantical information for the current key frame. Consequently, this might result in some incorrectness or a drastic fluctuation on speech or visual word topic trend estimation if using general time dependent topic models such as Seq-CI-HDP and Seq-cHDP. Fig. 4.11 and Fig. 4.12 show the results of topic trend estimation in video No.161 with Seq-CI-HDP and Seq-cHDP, respectively. In Fig. 4.11 with Seq-CI-HDP, although the topic trend estimation on visual words looked smooth and steady, when the number of speech words drastically decreased in Frame #37 and #38, the quality of the topic estimation for these speech words became worse, and some speech words were assigned to some wrong topics. This is because such weak

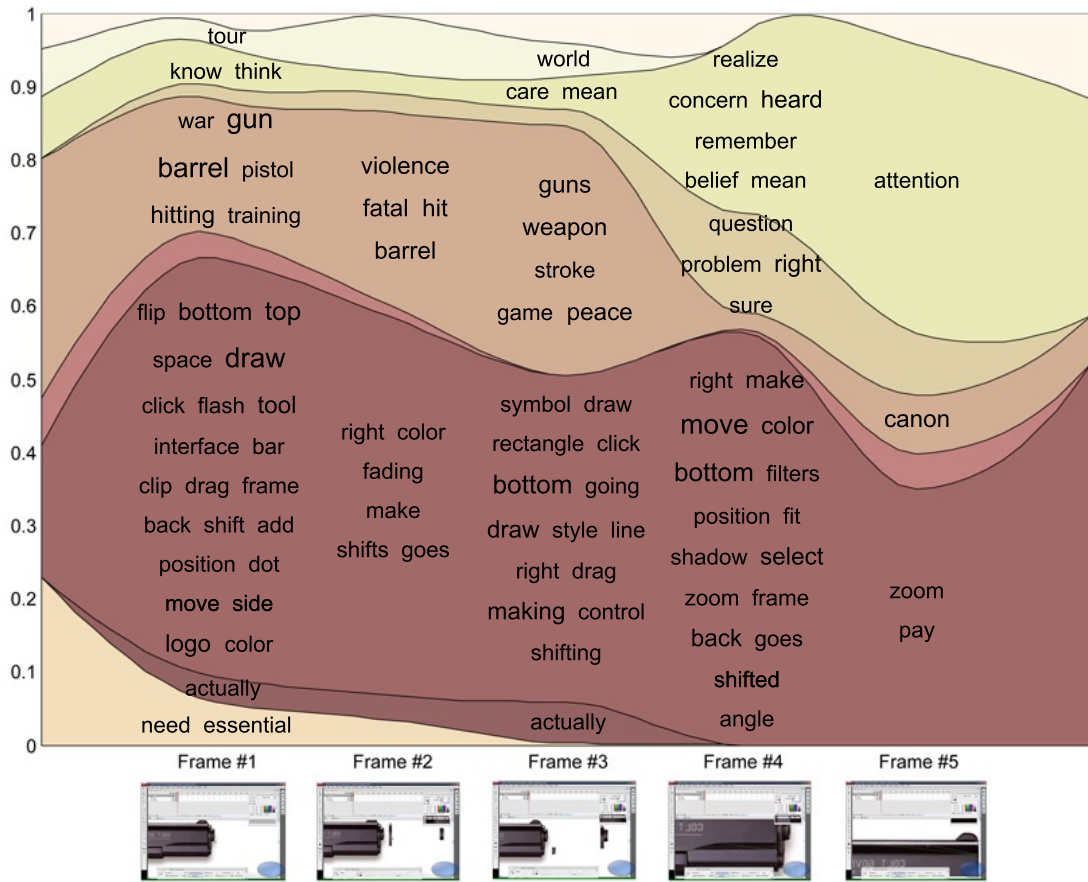


Fig. 4.10 Area graph for topic distribution in video No.984 (©AveLx) with Seq-Sym-cHDP, superimposed with speech words.

dependencies on multimodal modeling cannot fix such incorrectness on topic estimation caused by insufficient speech words, since all the modalities were conditionally independent for assigning topics to their own words in Seq-CI-HDP. In Fig. 4.12 with Seq-cHDP¹, the topic trend estimation on visual words showed a drastic fluctuation when the number of the pivot speech words drastically decreased in Frame #37 and #38, the effectiveness of topic trend estimation became unstable. This is because such unidirectional dependency on multimodal modeling made the topic learning for non-pivot visual words follow the learned topic distribution over pivot speech words, and the topic trend estimation for visual words was misled by a few pivot speech words in this scenario. However, according to Fig. 4.8, Seq-Sym-cHDP effectively fixed these problems caused by such imbalanced multimodal data, and improved the quality of topic trend estimation via balancing the pivot assignments

¹In this example, I predefined the pivot to speech data in Seq-cHDP. However, in the case of predefined the pivot to visual data in Seq-cHDP, the topic trend estimation was not influenced obviously, since visual data were dominant in the example of the video No.161.

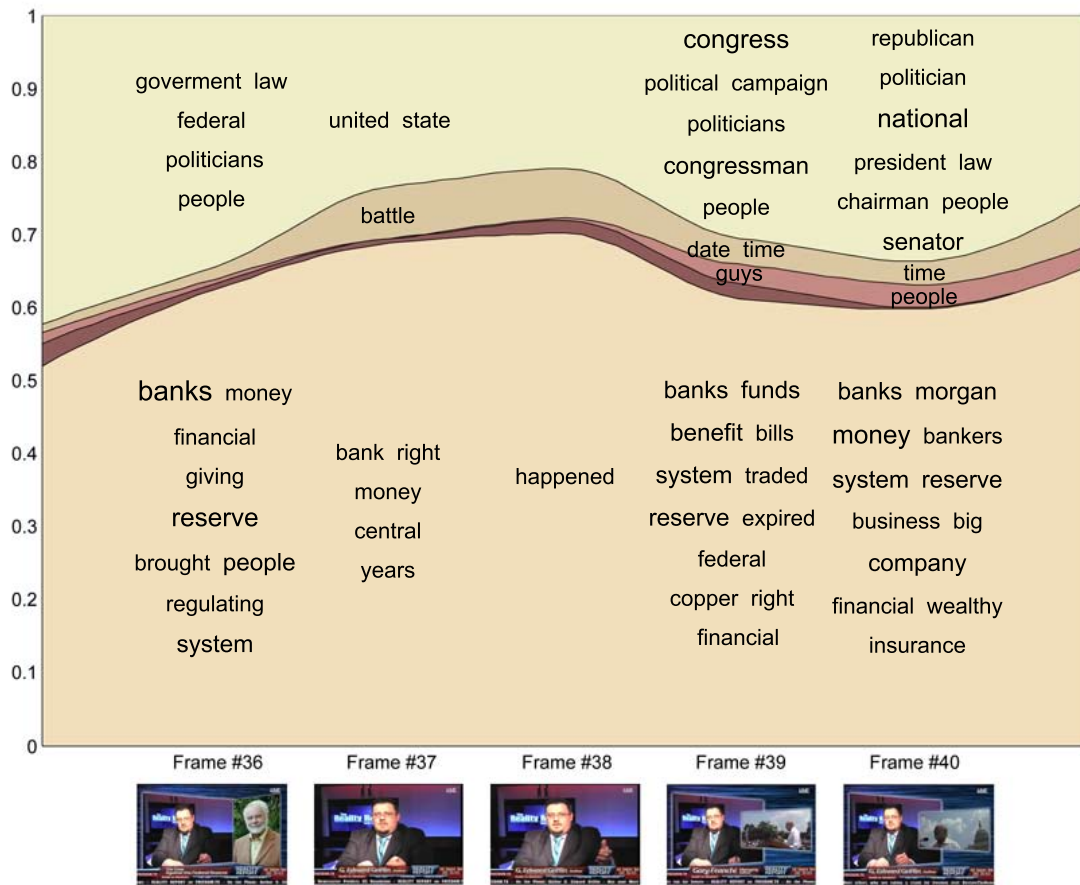


Fig. 4.11 Area graph for topic distribution in video No.161 (©The Reality Report) with Seq-CI-HDP, superimposed with speech words.

between speech words and visual words with the symmetric correspondence mechanism, as discussed in Section 4.3.3.

4.5.4 Quantitative Evaluation

The quantitative evaluation is based on evaluating the performance of the models with genre classification. In this task, there are totally three nonparametric topic models Seq-CI-HDP, Seq-cHDP and symmetric correspondence HDP (Sym-cHDP) adopted as baseline models. Note that Sym-cHDP is a simplified version of Seq-Sym-cHDP without the time-dependency mechanism. For Seq-cHDP, I compared two different scenarios by respectively predefining the pivot to visual data and speech data within Seq-cHDP. Similarly, I also performed the parameter tuning tasks for other baseline models beforehand, then the results showed that all the baseline models worked most effectively when setting the controlling parameters with

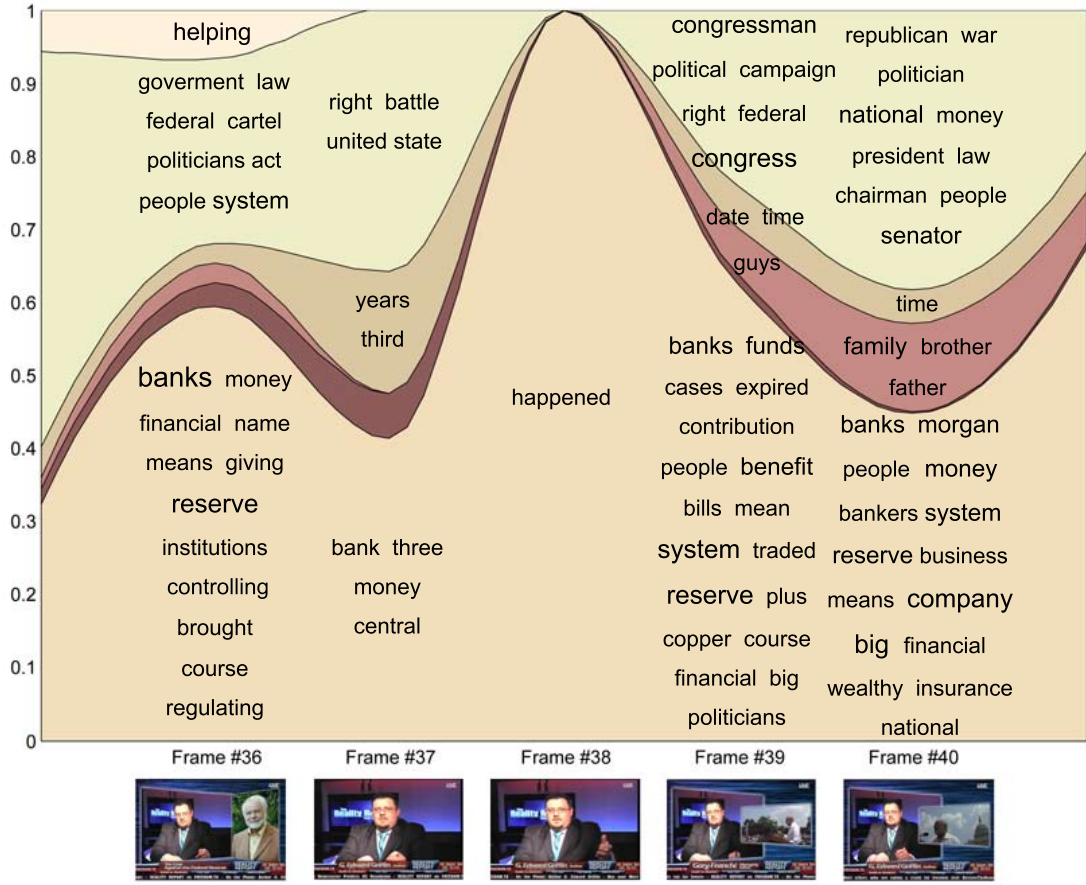


Fig. 4.12 Area graph for topic distribution in video No.161 (©The Reality Report) with Seq-cHDP, superimposed with speech words.

$w = 0.5^2$ and $\lambda = 1.0$, which are in accordance with the ones in Seq-Sym-cHDP, referred to Section 4.5.2.

Finally, the results of video genre classification for all the baseline models and Seq-Sym-cHDP are shown in Table 4.2, where Seq-Sym-cHDP outperformed the other baselines with both the results of accuracy (Micro-F1) and Macro-F1. This demonstrates that the symmetric correspondence mechanism incorporated in Seq-Sym-cHDP can make the model work more effectively in the case of video categorization. In addition, Sym-cHDP showed worse performance with more instability compared with other topic models, which demonstrates time-dependency mechanism incorporated in time dependent topic models plays a very important role in video data modeling as well.

²Note that there is no time-dependency mechanism in Sym-cHDP, therefore, w is permanently set to 0 in the case of Sym-cHDP during the experimental phase.

Table 4.2 Performance of Genre Classification.

	Accuracy	Macro-F1
Seq-Sym-cHDP	0.3621 ± 0.0136	0.0923 ± 0.0109
Seq-CI-HDP	0.3188 ± 0.0183	0.0678 ± 0.0143
Seq-cHDP(Speech)	0.3209 ± 0.0154	0.0675 ± 0.0122
Seq-cHDP(Visual)	0.3336 ± 0.0186	0.0708 ± 0.0146
Sym-cHDP	0.3036 ± 0.0291	0.0568 ± 0.0187

4.6 Summary

To further improve the multimodal data modeling method in Bayesian nonparametric topic models, I presented a sequential symmetric correspondence hierarchical Dirichlet processes (Seq-Sym-cHDP) model for video semantic analysis in this chapter.

Previously, I proposed a sequential correspondence HDP (Seq-cHDP) model based on the hierarchy of video data to analyze video semantical information in Chapter 3. For video multimodal data modeling, Seq-cHDP incorporates a unidirectional correspondence mechanism, which defines a pivot data type (modality), and then makes other non-pivot data types (modalities) generate their latent topics based on the topic distribution over pivot data. However, such correspondence mechanism makes Seq-cHDP lack flexibility, since the pivot data type needs to be predefined and fixed throughout the whole generative process. Then, I proposed another sequential conditionally independent HDP (Seq-CI-HDP) model, which assumes that all the types of data (modalities) are conditionally independent for generating their own data samples. However, the dependencies among all the multimodal data within Seq-CI-HDP are relatively weak.

Hence, I incorporate a more flexible symmetric correspondence mechanism into Seq-Sym-cHDP to improve the multimodal data modeling, which can spontaneously select the pivot data type when generating a topic for a corresponding word. Through this method, the dependencies among all the multimodal data can be flexibly balanced via managing the pivot flag assignment. Then, the Seq-Sym-cHDP model is inferred based on posterior representation samplers, and implemented with a cascaded Gibbs sampling method.

In experimentations, for qualitative evaluation, Seq-Sym-cHDP showed a significant flexibility on trend estimation for latent topics within the video even if when facing an extreme scenario such as extremely imbalanced multimodal data. For quantitative evaluation, Seq-Sym-cHDP achieved the best performance compared with other baselines on video genre classification.

However, Seq-Sym-cHDP is still an unsupervised Bayesian nonparametric topic model, which learns the latent topics from video data in an unsupervised learning manner. In Chapter 5, I will introduce supervised topic models for real applications.

Chapter 5

Supervised Bayesian Nonparametric Topic Modeling Methods for Multi-class Video Classification

5.1 Overview

Bayesian nonparametric topic models, such as hierarchical Dirichlet processes (HDP) [4], offer an effective way to learn latent properties from data. Based on original HDP model, I have respectively presented a sequential correspondence HDP (Seq-cHDP) model and a sequential symmetric correspondence HDP (Seq-Sym-cHDP) model for video semantic analysis in Chapter 3 and Chapter 4. However, all the above nonparametric topic models are unsupervised models, which cluster semantically or characteristically related features into a specific latent topic in an unsupervised learning manner, and such a learned topic is hard to interpret or map to a real class.

To tackle above problem, this chapter thus seeks a supervised learning method for multi-class video classification. Regarding the previous studies, McAuliffe et al. [16], for the first time, proposed a supervised topic model, *i.e.*, supervised latent Dirichlet allocation (sLDA), to regard document-level numerical rating as a regression response. For nonparametric topic models, Dai et al. [17] recently have developed such a supervised method in an HDP-based model, named supervised HDP (sHDP), which considers learned latent topics to be predictive of document responses. Inspired by these previous works, this chapter proposes a supervised sequential symmetric correspondence HDP (Sup-SSC-HDP) model for multi-class video classification, which can directly capture discrete video classes as classification response. As

a supervised learning version of Seq-Sym-cHDP, Sup-SSC-HDP is still a video data oriented nonparametric topic model that incorporates a time-dependency mechanism (for modeling the relationship among video frames) and a symmetric correspondence mechanism (for modeling multimodal data from videos) to offer a suitable learning for videos. A generalized linear model (GLM) is used to model the relationship between empirical topic frequencies over multimodal data and video class, and a softmax regression algorithm is used to learn their regression coefficients. During the testing phase, Sup-SSC-HDP computes video class likelihoods conditioned on trained regression coefficients, and finally predict a specific class for each test video.

The rest of this chapter is organized as follows. Section 5.2 firstly proposes a supervised model based on unimodal Seq-HDP. Then, a supervised sequential symmetric correspondence HDP (Sup-SSC-HDP) model extended from supervised Seq-HDP is proposed multi-class video classification in Section 5.3. Section 5.4 specifies inference method and video class prediction for Sup-SSC-HDP. Finally, experimental evaluations and summary are respectively presented in Section 5.5 and Section 5.6.

5.2 Supervised Sequential HDP

Although latent topics learned with unsupervised topic models are adequately expressive, they are not appropriate for a classification task, since the observed class information for each document is not exploited during training. Supervised topic models can overcome such problems by jointly learning topics and their regression coefficients for the document responses.

In Chapter 3, I firstly proposed an unsupervised sequential HDP (Seq-HDP) for handling the time dependencies between neighboring video frames within a video. In this section, I also start to develop a supervised method based on Seq-HDP.

To model the relationship between covariates and response, a generalized linear model (GLM) is widely used, which is still available even if the covariates and the response are non-linearly related or the response is not based on a Gaussian assumption. For the supervised Seq-HDP model, I thus use GLM to model the relationship between the empirical frequencies of the topics in the f -th video $\bar{\mathbf{z}}_f$ and the video response y_f , where I define a linear predictor $\boldsymbol{\varpi} = \boldsymbol{\eta}^\top \bar{\mathbf{z}}_f$, $\boldsymbol{\eta}$ denotes regression coefficients corresponding to the empirical frequencies of the topics $\bar{\mathbf{z}}_f$, and I consider that the video response y_f obeys an exponential family distribution. Then, a canonical link function $g^{-1}(\cdot)$ is used to relate $\boldsymbol{\varpi}$ to the mean $E[y_f|\cdot]$ of y_f , which is expressed as $E[y_f|\cdot] = g^{-1}(\boldsymbol{\varpi})$.

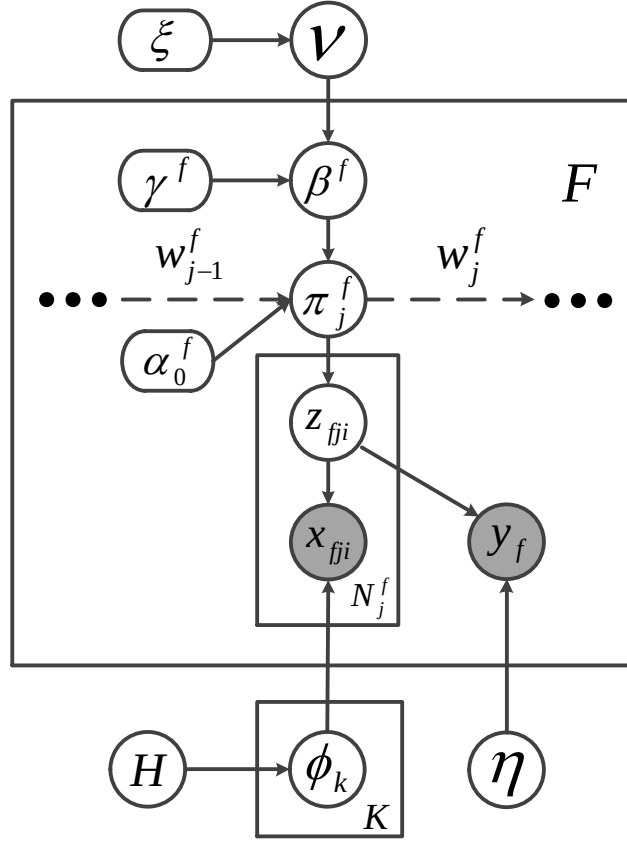


Fig. 5.1 Graphical model representation of supervised Seq-HDP.

A simplified graphical model representation of the stick-breaking construction of supervised Seq-HDP is depicted in Fig. 5.1. Its generative process is as follows.

- **Step 1:** Successively draw component weight variables ν , β^f , and π_j^f based on following procedures,

$$\begin{aligned} \nu &\sim GEM(\xi) \\ \beta^f &\sim DP(\gamma^f, \nu) \\ \pi_j^f &\sim DP(\alpha_0^f, \pi_{j-1}^f) \end{aligned} \quad (5.1)$$

where $GEM()$ represents a GEM process [19], and $DP()$ represents a Dirichlet process. ξ , γ^f and α_0^f are corresponding hyperparameters. Note that the generation of π_j^f is influenced by both β^f and π_{j-1}^f because of the time dependency. Hence, π_j^f is expressed as $\pi_j^f = w_{j-1}^f \pi_{j-1}^f + (1 - w_{j-1}^f) \beta^f$.

- **Step 2:** For each word, draw its topic z_{fji} from π_j^f , then sample the word x_{fji} conditioned on both sampled z_{fji} and ϕ_k .

• **Step 3:** For each video, draw the video response with $y_f \sim \text{ExpFam}(g^{-1}(\eta^\top \bar{\mathbf{z}}_f), \rho)$, where $\text{ExpFam}()$ is a distribution from the exponential family, and ρ is the dispersion parameter of the distribution.

For GLM, the distribution of response can have different specific forms due to the different applications, such as Gaussian, Bernoulli or multinomial. Here, I focus on multi-class video classification and thus model y_f as being distributed according to a multinomial distribution (the number of trials $n = 1$), and $y_f \in \{1, 2, \dots, R\}$, where R is the total number of classes. Because of GLM, I express this distribution with an exponential family representation:

$$P(y_f = c | \eta, \bar{\mathbf{z}}_f) = \frac{\exp(\eta^c \bar{\mathbf{z}}_f)}{\sum_{r=1}^R \exp(\eta^r \bar{\mathbf{z}}_f)} \quad (5.2)$$

where η is an $R \times K(\infty)$ regression coefficient matrix, and η^c is a coefficient vector extracted from the c -th row of η . Note that this multi-class supervised learning process is also known as softmax regression.

5.3 Supervised Sequential Symmetric Correspondence HDP

To model the video multimodal data¹ with more flexibility, in Chapter 4, I previously proposed a sequential symmetric correspondence HDP (Seq-Sym-cHDP) model, where the pivot data type among multimodal data is spontaneously selected by using a pivot flag generator when generating each topic. In this section, I incorporate this symmetric correspondence mechanism into the supervised Seq-HDP framework, and then devise a new supervised sequential symmetric correspondence HDP (Sup-SSC-HDP) model for multi-class video classification.

A simplified graphical model representation of Sup-SSC-HDP is shown in Fig. 5.2, where I use l to represent the index of the data type (modality), and $l \in \{1, \dots, L\}$, and $N_{fj}^{(l)}$ is the total count for the words of data type l in frame j of video f , and all the $\{\phi_k^{(l)}\}_{l=1}^L$ drawn from H share the same topic space. The detailed generative process of Sup-SSC-HDP is as follows.

- **Step 1:** Successively draw component weight variables $\mathbf{v}$, β^f , and π_j^f based on Eq. 5.1.
- **Step 2:** Draw a multinomial pivot flag generator σ_{fj} from $\text{Dir}(\lambda)$.

¹In previous chapters on video semantic analysis, two types of data were drawn from videos, and respectively converted into visual words and speech words, which were used in proposed nonparametric topic models.

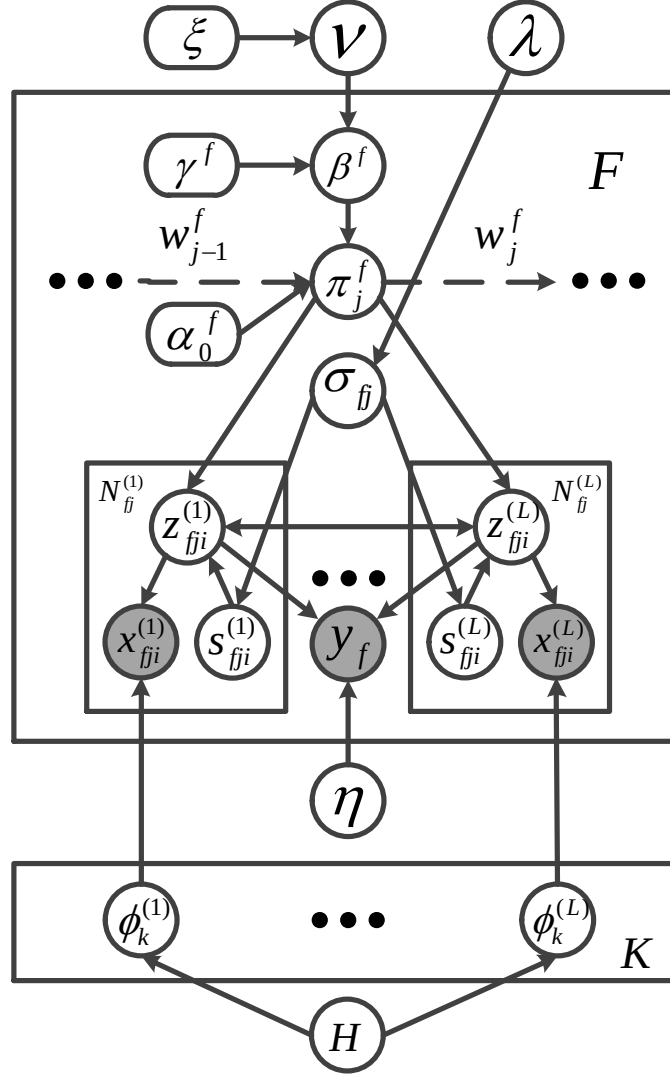


Fig. 5.2 Graphical model representation of Sup-SSC-HDP.

- **Step 3-a:** For each word of data type l in frame j of video f , draw a pivot flag $s_{fji}^{(l)}$ from $Multinomial(\sigma_{fj})$.
- **Step 3-b:** If $s_{fji}^{(l)} = l$, draw a topic $z_{fji}^{(l)}$ from π_j^f . Otherwise if $s_{fji}^{(l)} = h \neq l$, draw a topic $z_{fji}^{(l)}$ from $Uniform(z_{fj1}^{(h)}, \dots, z_{fjH_{fj}}^{(h)})$.
- **Step 3-c:** Conditioned on sampled $z_{fji}^{(l)}$, draw a word $x_{fji}^{(l)}$ from $f(x_{fji}^{(l)} | \phi_k^{(l)}, k = z_{fji}^{(l)})$.
- **Step 4:** For each video, draw a video class y_f from a distribution $P(y_f | \eta, \bar{z}_f)$.

Here, $Dir(\lambda)$ indicates a Dirichlet distribution with a hyperparameter λ , $Multinomial()$ indicates a Multinomial distribution, and $Uniform()$ indicates a uniform distribution. $f(x_{fji}^{(l)} | \phi_k^{(l)})$

indicates a word distribution conditioned on $\phi_k^{(l)}$. η is the regression coefficient, and $\bar{\mathbf{z}}_f$ is the empirical frequencies of the topics within video f .

Similar to the generative process of Seq-Sym-cHDP, the way of generating the topic is also chosen by a sampled pivot flag, which determines the current pivot data type. Its detailed explanation can be found in Section 4.3.3 of Chapter 4. Different from $N_{fj}^{(h)}$, $H_{fj}^{(h)}$ represents the number of words that have already been assigned topics in the current step, which means $H_{fj}^{(h)} \leq N_{fj}^{(h)}$. Through this method, the generative processes for different multimodal data can be modeled in a symmetric way, and the dependencies among all the multimodal data can be appropriately taken into account.

5.4 Inference Method and Class Prediction

Firstly, on the basis of posterior representation sampler [20], I develop a inference scheme for inferring the Sup-SSC-HDP. Then, I introduce a softmax regression method to update the regression coefficients η , and finally derive an approach to predict a video class for each test video.

5.4.1 Posterior Representation

Since the generative process of all the component weights in Sup-SSC-HDP are completely identical to the one in Seq-Sym-cHDP, here I recall the sampling method for all the component weights from Section 4.4.1 of Chapter 4.

Due to the posterior representation sampler [20], every original infinite component weight vector is reformulated with a new augmentable finite vector that consists of K components and a promising component u . Below, I give the new representation of $\mathbf{v}$ and its sampling procedure:

$$\mathbf{v} = (v_1, \dots, v_K, v_u) \sim \text{Dir}(M_1, \dots, M_K, \xi) \quad (5.3)$$

where a new variable M_k^f used for sampling $\mathbf{v}$ indicates the metatable counts in the Chinese restaurant franchise (CRF) [13] process of three-layer HDP, which is explained in Section 3.4.1 of Chapter 3. M_k is the marginal form of M_k^f (i.e., $M_k = \sum_f M_k^f$).

Similarly, β^f can be represented and sampled using:

$$\beta^f = (\beta_1^f, \dots, \beta_K^f, \beta_u^f) \sim \text{Dir}(\tilde{\beta}_1^f, \dots, \tilde{\beta}_K^f, \tilde{\beta}_u^f) \quad (5.4)$$

where $\tilde{\beta}_k^f = \gamma^f v_k + T_{fk}^{0 \rightarrow \cdot}$ and $\tilde{\beta}_u^f = \gamma^f v_u$. The variable $T_{fk}^{0 \rightarrow j}$ indicates the table counts of CRF yielded by β_k^f in the video frame j , and $T_{fk}^{0 \rightarrow \cdot}$ is the marginal form of $T_{fk}^{0 \rightarrow j}$ (i.e., $T_{fk}^{0 \rightarrow \cdot} = \sum_j T_{fk}^{0 \rightarrow j}$).

For the local component weights, π_j^f can be represented and sampled as:

$$\pi_j^f = (\pi_{j1}^f, \dots, \pi_{jK}^f, \pi_{ju}^f) \sim \text{Dir}(\tilde{\pi}_{j1}^f, \dots, \tilde{\pi}_{jK}^f, \tilde{\pi}_{ju}^f) \quad (5.5)$$

where two estimators $\tilde{\pi}_{jk}^f$ and $\tilde{\pi}_{ju}^f$ can be linearly formulated with both β_k^f (or β_u^f) and $\pi_{(j-1)k}^f$ (or $\pi_{(j-1)u}^f$) because of the time-dependency mechanism, as follows:

$$\begin{aligned} \tilde{\pi}_{jk}^f &= \alpha_0^f w_{j-1}^f \pi_{(j-1)k}^f + \alpha_0^f (1 - w_{j-1}^f) \beta_k^f + C_{fjlk}^{(\cdot)} + T_{fk}^{j \rightarrow j+1} \\ \tilde{\pi}_{ju}^f &= \alpha_0^f w_{j-1}^f \pi_{(j-1)u}^f + \alpha_0^f (1 - w_{j-1}^f) \beta_u^f \end{aligned} \quad (5.6)$$

where $C_{fjlk}^{(l)}$ indicates the counts for the word with data type l in frame j of video f possessing the pivot flag with the same data type l when the sampled topic is k , and $C_{fjlk}^{(\cdot)} = \sum_{(l)} C_{fjlk}^{(l)}$. $T_{fk}^{j \rightarrow j+1}$ indicates another table counts of CRF that occurs because of the time-dependency mechanism (from π_{jk}^f to $\pi_{(j+1)k}^f$). Note that the details of the sampling method for the metatable counts M_k^f , table counts $T_{fk}^{j-1 \rightarrow j}$ and $T_{fk}^{0 \rightarrow j}$, can be found in Section 3.4.2 of Chapter 3.

Then, a full conditional joint likelihood is derived for sampling both the topic $z_{fji}^{(l)}$ and pivot flag $s_{fji}^{(l)}$ for each word.

$$P(z_{fji}^{(l)} = k, s_{fji}^{(l)} = m | x_{fji}^{(l)}, \lambda, \eta, \dots) \propto P(s_{fji}^{(l)} = m | \lambda) P(z_{fji}^{(l)} = k | \cdot) P(x_{fji}^{(l)} | \phi_k^{(l)}) P(y_{fj} | z_{fji}^{(l)} = k, \eta) \quad (5.7)$$

I describe these expanded likelihoods in the following. $P(s_{fji}^{(l)} = m | \lambda)$ indicates the pivot flag likelihood, which is formulated as:

$$P(s_{fji}^{(l)} = m | \lambda) = \frac{C_{fjm}^{-fjli} + \lambda}{C_{fjm}^{-fjli} + \sum_{m' \neq m} C_{fjm'}^{-fjli} + L\lambda} \quad (5.8)$$

where C_{fjm} indicates the counts for the pivot flag pointing to data type m over all the multimodal data in frame j of video f , and the superscript $-fjli$ for a variable indicates the terms excluding the $fjli$ th term.

$P(z_{fji}^{(l)} = k | \cdot)$ indicates the conditional likelihood for topic k , if $s_{fji}^{(l)} = l$, $P(z_{fji}^{(l)} = k | \pi_j^f)$ is equal to π_{jk}^f (or π_{ju}^f when k is newborn), if $s_{fji}^{(l)} = h \neq l$, there is $P(z_{fji}^{(l)} = k | Z_{fj}^{(h)}) = \frac{n_{fjk}^{(h)}}{N_{fj}^{(h)}}$ where $Z_{fj}^{(h)}$ is a vector that records all the topic assignments for the words of data type h in frame j of video f , and $n_{fjk}^{(h)}$ is the count for the assigned topic k in the same domain.

$P(x_{fji}^{(l)} | \phi_k^{(l)})$ indicates the conditional word likelihood, if the topic k is previously used, there is:

$$P(x_{fji}^{(l)} | \phi_k^{(l)}) = \frac{n_{kv^{(l)}}^{-fjli} + \tau}{\sum_{v^{(l)}} n_{kv^{(l)}}^{-fjli} + V^{(l)} \tau} \quad (5.9)$$

if k is a newborn topic, the likelihood is $1/V^{(l)}$, where $v^{(l)}$ is the index of words in terms of the vocabulary in data type l , $n_{kv^{(l)}}$ is the counts for the word $v^{(l)}$ assigned to topic k , $V^{(l)}$ is the vocabulary size of data type l , τ is a hyperparameter for a Dirichlet distribution, and $H = \text{Dir}(\tau)$.

$P(y_{fj} | z_{fji}^{(l)} = k, \eta)$ indicates the conditional class likelihood, which is formulated by:

$$P(y_{fj} = c | z_{fji}^{(l)} = k, \eta) = \frac{\exp(\sum_{k'} \eta_{k'}^c \bar{z}_{fjk'}^{-fjli} + \frac{\eta_k^c}{\sum_l N_{fj}^{(l)}})}{\sum_{r=1}^R \exp(\sum_{k'} \eta_{k'}^r \bar{z}_{fjk'}^{-fjli} + \frac{\eta_k^r}{\sum_l N_{fj}^{(l)}})} \quad (5.10)$$

where I assume $y_{fj} = y_f$ during training. If k is newborn, I reformulate each η^r with $\eta^r = (\eta^r, \eta_{k_{new}}^r)$, where $\eta_{k_{new}}^r$ is drawn from a Gaussian prior.

The hyperparameters ξ , γ^f and α_0^f corresponding to ν , β^f and π_j^f are sampled from a gamma prior. The hyperparameter λ is another controlling parameter that will be discussed in the experimental section, and the hyperparameter τ is a constant parameter in this paper. During training, the regression coefficients η can be updated with a gradient descent algorithm, which will be described in Section 5.4.3.

5.4.2 Cascaded Gibbs Sampling

Finally, I derive a cascaded Gibbs sampling method for implementing the inference scheme of Sup-SSC-HDP based on the posterior representation sampler, which is summarized as Algorithm 4.

In the initialization phase, words are randomly allocated to pivot flags and topics, and the regression coefficients η are drawn from a Gaussian prior. During the Gibbs sampling phase, the total number of topics K increases if the promising component is triggered to sample a

Algorithm 4: Cascaded Gibbs Sampling for Inferring Sup-SSC-HDP

Initialization;
for ($f = 1; f \leq F; f++$) **do**
 for ($j = J^f; j \geq 1; j--$) **do**
 for ($l = 1; l \leq L; l++$) **do**
 Sample pivot flags $S_{fj}^{(l)}$ and topics $Z_{fj}^{(l)}$;
 end
 for ($k = 1; k \leq K; k++$) **do**
 Sample table counts $T_{jk}^f, T_{fk}^{j-1 \rightarrow j}$ and $T_{fk}^{0 \rightarrow j}$;
 end
 end
 for ($k = 1; k \leq K; k++$) **do**
 Sample metatable counts M_k^f ;
 end
end
Update regression coefficients η ;
Sample hyperparameters ξ, γ^f and α_0^f ;
Sample overall component weights v ;
for ($f = 1; f \leq F; f++$) **do**
 Sample global component weights β^f ;
 for ($j = 1; j \leq J^f; j++$) **do**
 Sample local component weights π_j^f ;
 end
end

newborn topic when sampling $z_{fji}^{(l)}$. All the component weight parameters are thus required to augment their components with this newborn topic in the sampler [4]. In the testing phase, the posterior of $\bar{\mathbf{z}}_f$ is necessary over the test videos for class prediction. Hence, I remove the $P(y_{fj}|\cdot)$ term from the full conditional joint likelihood for sampling $z_{fji}^{(l)}$ and $s_{fji}^{(l)}$ in the testing period, which makes that the inference on the test videos is identical to unsupervised learning.

5.4.3 Softmax Regression and Class Prediction

During the training period, the regression coefficients η need to be updated for each iteration (or a number of iterations) of cascaded Gibbs sampling system. As mentioned in Section 5.2, this can be considered a softmax regression problem because of the multi-class video classification task.

Based on the learned $\bar{\mathbf{z}}_f$ and observed y_f for each video, the cost function is expressed as:

$$J(\eta) = -\frac{1}{F} \left[\sum_{f=1}^F \sum_{r=1}^R 1\{y_f = r\} \log P(y_f = r | \eta, \bar{\mathbf{z}}_f) \right] + \frac{\varepsilon}{2} \sum_{r=1}^R \eta^{r\top} \eta^r \quad (5.11)$$

where F is the total number of videos, $1\{\cdot\}$ is an indicator function that takes on a value of 1 if its argument is true. $P(y_f = r | \eta, \bar{\mathbf{z}}_f)$ is the class probability defined in Eq. 5.2. Note that I add a weight decay term $\frac{\varepsilon}{2} \sum_{r=1}^R \eta^{r\top} \eta^r$ in this cost function to make $J(\eta)$ strictly convex, and $\varepsilon > 0$. To minimize the $J(\eta)$, I use a gradient descent algorithm and then find the optimized η . The gradient for each η^r is expressed as follows.

$$\nabla_{\eta^r} J(\eta) = -\frac{1}{F} \sum_{f=1}^F \left[\bar{\mathbf{z}}_f (1\{y_f = r\} - P(y_f = r | \eta, \bar{\mathbf{z}}_f)) \right] + \varepsilon \eta^r \quad (5.12)$$

During testing, the empirical frequencies of topics $\bar{\mathbf{z}}_f$ for each video are computed over a number of iterations. I average these $\bar{\mathbf{z}}_f$, and obtain $E[\bar{\mathbf{z}}_f]$. Based on the $E[\bar{\mathbf{z}}_f]$ and the η learned from training, I predict a class $\tilde{c}_f$ for each test video as follows.

$$\tilde{c}_f = \arg \max_{c \in (1, \dots, R)} \frac{\exp(\eta^{c\top} E[\bar{\mathbf{z}}_f])}{\sum_{r=1}^R \exp(\eta^{r\top} E[\bar{\mathbf{z}}_f])} \quad (5.13)$$

5.5 Experiments

5.5.1 Experimental Setup

For evaluating the performance of Sup-SSC-HDP on multi-class video classification, I also used a *test_data* dataset extracted from *MediaEval-2011 Genre Tagging Task* [22] to conduct this experiment. In this dataset, two types of data were drawn from videos. For each video, a sequence of key frames were extracted as image data, meanwhile a number of speech transcript words considered to be speech data were allocated to each key frame via an automatic speech recognition (ASR) [22]. Because of the topic models, these image data and speech data need to be processed to independent visual words and speech words, respectively. The details on extracting visual words and speech words can be found in Section 3.5.1 of Chapter 3. The summary of the experimental dataset is given in Table 5.1.

During the experiment, four different multimodal topic models, *i.e.*, Seq-Sym-cHDP, sequential conditionally independent HDP (Seq-CI-HDP), Seq-cHDP and symmetric cor-

Table 5.1 Summary of Experimental Dataset.

the number of video files	1726
the number of genre labels	26
the number of key frames	40588
the number of visual words	166803k
the number of speech words	1318k
the vocabulary size of visual words	1000
the vocabulary size of speech words	17596

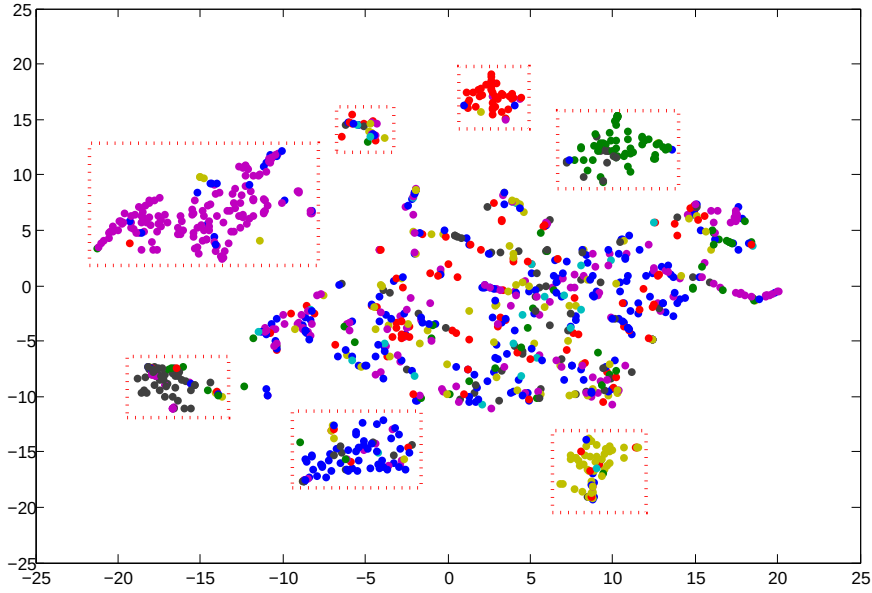


Fig. 5.3 t-SNE 2D embedding of topic frequencies with Sup-SSC-HDP.

respondence HDP (Sym-cHDP) were chosen as baseline models². For the initializations on these topic models, the hyperparameters ξ , γ^f , and α_0^f were sampled from a gamma prior $\text{Gamma}(1.0, 1.0)$. Referring to the previous chapters, I set all the transfer weights w_j^f to a constant, 0.5. In addition, τ was set to 1.0. The controlling parameter λ was tuned with different values, which is described in Section 5.5.3. The CGS system totally ran 1000 iterations for each training or testing task so as to let all the variables in the model fully converge.

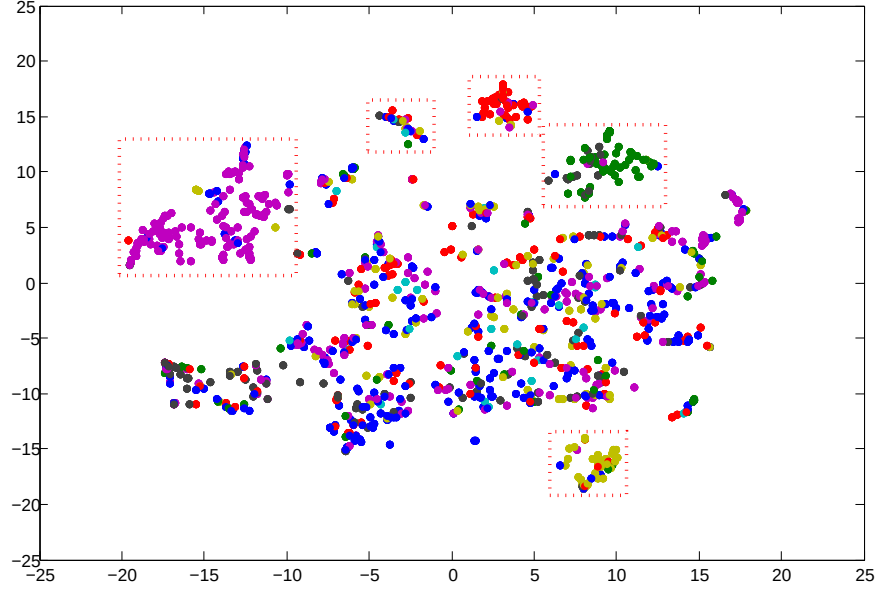


Fig. 5.4 t-SNE 2D embedding of topic frequencies with Seq-Sym-cHDP.

5.5.2 Qualitative Evaluation

To demonstrate intuitively how Sup-SSC-HDP performed in the experiments, I utilized the t-SNE stochastic neighborhood embedding [28] method to depict the 2D embedding of the learned topic frequencies $\bar{\mathbf{z}}_f$ by Sup-SSC-HDP and unsupervised Seq-Sym-cHDP, respectively, as shown in Fig. 5.3 and Fig. 5.4, where each dot stands for a video, and each type of color superimposed on the dot stands for a specific class for videos, each box plotted with a red dotted line stands for a visually observed cluster. In this figure, clustering phenomenon with respect to dots of the same color occurred in both 2D embedding graphs, which demonstrate that some of the videos within the same class were successfully grouped after learning via Sup-SSC-HDP and Seq-Sym-cHDP. Furthermore, Sup-SSC-HDP performed better than Seq-Sym-cHDP in clustering and separating the videos into different classes, which means that the supervised modeling method learns better how to map topics to different classes. However, I also noticed that there a number of videos were not well grouped or were grouped into incorrect clusters. The causes of these failures are explained below.

1) The video counts for different classes are imbalanced in this dataset, some of the major classes have more than 100 videos per class, while some of the minor classes have less than 20 videos per class. This imbalance especially affects the learning for videos in the minor class.

²All these four different multimodal topic models are specified in Chapter 4.

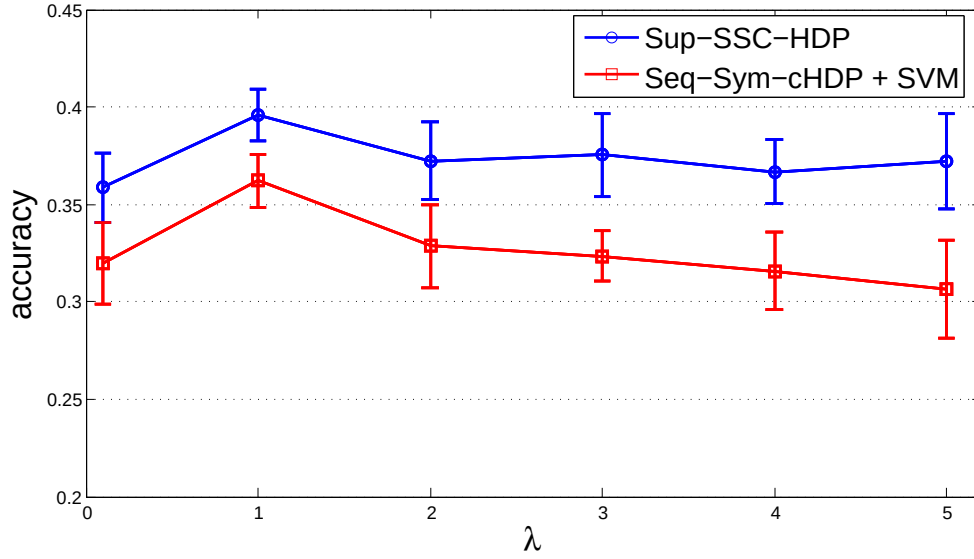


Fig. 5.5 Performance of video classification with different λ .

2) The ranges of definition for some of the classes overlap. For instance, two classes *science* and *educational* exist in our dataset, however, sometimes, scientific video can be regarded as educational video. Such ambiguous cases are problematic for video learning.

5.5.3 Quantitative Evaluation

Here, I focus on evaluating the performance of multi-class video classification with topic models. To comprehensively assess our model, the comparisons were made of: (1) supervised and unsupervised methods, (2) symmetric correspondence and other multimodal modeling methods, and (3) sequential and nonsequential methods. Note that the unsupervised topic model cannot directly classify videos, so I needed to input its learned topic frequencies into an extra classifier (such as SVM) to conduct the classification. In addition, I used a five-fold cross-validation scheme in the experiments.

The hyperparameter λ in Sup-SSC-HDP is a controlling parameter that influences the generation of pivot flags. Here, I was interested in examining the performance of classification on Sup-SSC-HDP while varying λ . The results for Sup-SSC-HDP and an unsupervised version of Sup-SSC-HDP (*i.e.*, Seq-Sym-cHDP) with λ manually set at $\{0.1, 1, 2, 3, 4, 5\}$ in Fig. 5.5, and find that the models performed the best when $\lambda = 1$. In addition, Sup-SSC-HDP outperformed Seq-Sym-cHDP in each experiment of selecting λ .

Fig. 5.6 shows a comprehensive comparison of Sup-SSC-HDP and the baseline models, where Sym-cHDP is a nonsequential version of Seq-Sym-cHDP (*i.e.*, $w_j^f = 0$), and Seq-CI-

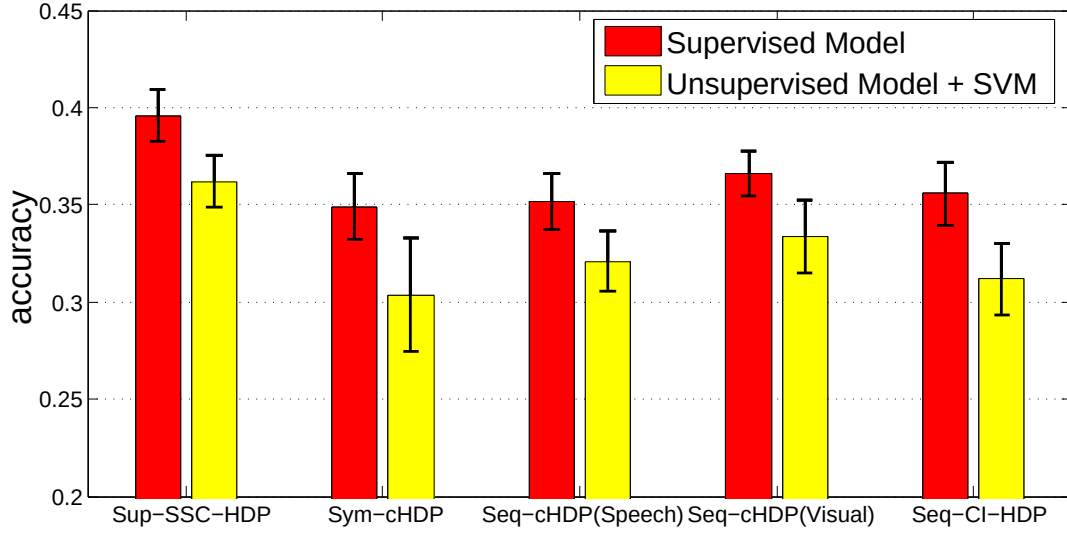
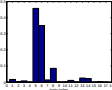
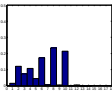
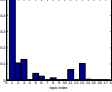
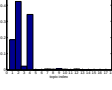


Fig. 5.6 Video classification performance of Sup-SSC-HDP and baseline models.

HDP is a non-correspondence version of Seq-cHDP or Seq-Sym-cHDP. Seq-cHDP had two forms, one setting speech data as the pivot data type, the other setting visual data as the pivot data type. According to the results, supervised method outperformed the unsupervised one in terms of classification accuracy for each HDP-based model, which demonstrates that learning the topics in such a supervised manner (relating topics to observed classes in training) improves the accuracy of classifying videos. Compared with other supervised baseline models, Sup-SSC-HDP achieved the best performance with the accuracy of 0.396 ± 0.013 . I thus infer that the time-dependency mechanism (in contrast to Sym-cHDP) and the symmetric correspondence mechanism (in contrast to Seq-cHDP and Seq-CI-HDP) incorporated in Sup-SSC-HDP contribute to a more precise video classification, as they can offer the best suitable modeling solution for video data.

Four video classification examples (three successes and one failure) are listed in Table 5.2. Here, Sup-SSC-HDP predicted the correct classes for videos expressing class-related visual and speech content, and the learned topic frequencies behaved discriminatively as well. However, the videos were sometimes hard to classify when their content was vague. The failure in Table 5.2, *i.e.*, the video titled “*How Data Can Improve Health Care*”, belongs to the class of *technology*, whereas its speech words include lots of *health*-related words, this made Sup-SSC-HDP misclassify this video as a *health* video. This problem is also discussed in Section 5.5.2 about the reasons of incorrect video clustering.

Table 5.2 Example Results on Video Classification with Sup-SSC-HDP.

	Video Title	Key Frames	Main Speech Words	Topic Frequency	True Class	Predicted Class
Successful Cases	Insight With Tariq Ali On Pakistan (©Frontline Club)		war, political, military, weapons, Pakistan, power, army, government, national, country...		politics	politics
	Web Expo - Things That Matter (©O'Reilly Media)		developers, technology, work, eBay, ecommerce, iPhone, online, PayPal, platform, Internet...		technology	technology
	Why We Don't Watch Soccer (©That Sports Show)		sports, game, teams, soccer, play, cup, matches, run, lead, catch...		sports	sports
Failure Case	How Data Can Improve Health Care (©O'Reilly Media)		health, medical, care, system, improve, measures, physicians, study, technical, hospital...		technology	health

5.6 Summary

This chapter presented a supervised sequential symmetric correspondence HDP (Sup-SSC-HDP) model for multi-class video classification, where a generalized linear model (GLM) was used to model the relationship between empirical topic frequencies over multimodal video data and video class. I also derived posterior representation samplers to infer the Sup-SSC-HDP, and implemented it by using a cascaded Gibbs sampling method. During the training phase, the regression coefficients corresponding to empirical topic frequencies are updated via softmax regression for each iteration of Gibbs sampling method.

In this work, two major contributions were made on Sup-SSC-HDP: (1) a supervised Bayesian nonparametric topic model, for the first time, was proposed for modeling multimodal video data in a supervised learning manner, (2) such supervised method incorporated in Sup-SSC-HDP effectively improved the accuracy of multi-class video classification compared with the performance of corresponding unsupervised nonparametric topic models proposed in the previous chapters.

In future works, I will consider to try alternative approach to update the regression coefficients in a Bayesian inference manner, such as slice sampling. Furthermore, a more effective supervised learning method such as maximum entropy discrimination [29] is expected to be incorporated in the nonparametric topic models to further improve the performance of video classification.

Chapter 6

Bayesian Nonparametric Multimodal Data Modeling Framework for Video Emotion Recognition

6.1 Overview

With the rapid development of computer vision and deep learning techniques, researchers are not merely content with some studies on simple video semantic analysis. Recently, they have been paying more attentions on video deep understanding, such as video emotion recognition, which is considered to be a very promising technique that can help us understand the emotion of people shown in a video via mining latent sentimental information from video features. However, learning emotions from video data is a quite challenging work. First of all, human sentiment is complex and diverse based on the theory about close interaction between cognitive processes and emotional appraisals [30]. Secondly, the expressions of emotion are sparsely distributed in a video. Thirdly, the multimodal data (such as image data and audio data extracted from video) processing and modeling for video emotion recognition are tricky as well.

Traditional methods for video emotion recognition on previous studies [31–34] mostly focused on extracting and learning relative sentimental patterns from low-level hand-crafted video features, which always have a bottleneck with respect to the performance of emotion recognition. Most recently, since the technique of convolutional neural networks (CNNs) [35–38] made a significant breakthrough on large-scale image classification tasks, researchers attempted to utilize such deep learning approach to further improve the performance of

emotion recognition for images or videos [39–41]. Although directly using such CNNs for image or video emotion recognition cannot perform as good as they did on image classification, these attempts were still meaningful, since they found important clues that the features extracted from high-level (deeper) layers in CNNs potentially involve emotional information [41].

In this chapter, a Bayesian nonparametric multimodal data modeling framework is proposed to analyze the video emotion based on extracted CNN-based deep features and MFCC features from videos. In this framework, a symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) model [42] is applied to mine the latent emotional events (topics) from both CNN-based deep features and MFCC features within a video. Through a series of comparative experiments, I can validate which CNN architecture is more appropriate for emotion recognition, and which layer in CNNs may contain more emotional features. Furthermore, the performance of the proposed framework for video emotion classification is also experimentally evaluated compared with several baseline methods.

The rest of this chapter is organized as follows. Section 6.2 gives a overview of Bayesian nonparametric multimodal data modeling framework. Section 6.3 introduces the CNN-based deep features used in this framework, and how to convert deep features and MFCC features into visual words and audio words, respectively. Section 6.4 introduces the Sym-cHDP model for mining the latent emotional events (topics) from both visual words and audio words within a video. In the end, comparative experiments and summary are respectively presented in Section 6.5 and Section 6.6.

6.2 Framework

Human emotion recognition for videos is a quite challenging work because of its complexity, diversity and sparsity. In this chapter, I develop a Bayesian nonparametric multimodal data modeling framework to learn the latent sentimental information from video data.

Similar to the task of video semantical analysis, before starting video emotion recognition, it is very necessary to determine what kind of features extracted from videos are appropriate for sentiment mining. Thanks to the rapid development of convolutional neural networks (CNNs) [35–38], some researchers believe that the image features extracted from high-level (deeper) layers in CNNs potentially contain some sentimental information according to the latest studies [41]. Moreover, Mel-frequency cepstral coefficient (MFCC) as one of the audio features was usually utilized for human emotion analysis in some previous studies [33, 41]. I realize that both image deep features and audio MFCC features can be simultaneously drawn

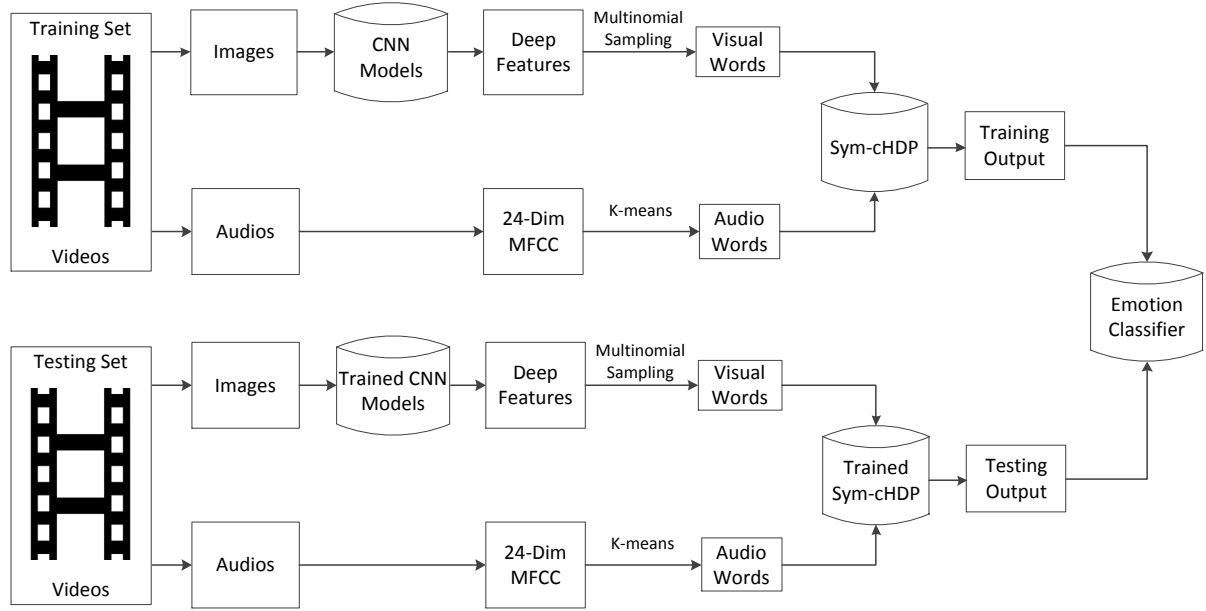


Fig. 6.1 A flowchart of framework for video emotion classification.

from videos, which inspires me to jointly use such two types of features for video emotion recognition.

Fig. 6.1 illustrates a flowchart of the proposed framework, which incorporates two core models: (1) a CNN model that deeply learns the images (video frames) extracted from videos, (2) a symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) model [42] that mines their latent emotional events (topics) from deep features and MFCC features. In the beginning, I respectively extract raw image data and raw audio data from each video. For images, I then input them to a CNN model. Generally, CNN architecture comprising a sequence of deep learning layers is straightforwardly applied for image classification [35–38]. In this framework, I only draw the deeply learned image deep features via the CNN model. The detailed description for CNN models and deep features will be specified in Section 6.3.1. Meanwhile, for audio data, a 24-dimensional MFCC descriptor is sampled over every 5 ms time-window with 50% overlap from entire soundtrack of each video. Such MFCC features are good audio representation for the case of emotion recognition.

In Chapter 4, I proposed a sequential symmetric correspondence HDP (Seq-Sym-cHDP) model for video semantic analysis. For simplicity, a two-layer version of Seq-Sym-cHDP, *i.e.*, Sym-cHDP model [42] is utilized to simultaneously deal with deep features and MFCC features in this framework. Similar to the case of semantical expression in videos, the structure of sentimental expression in video multimodal data has a similar form. For a video, the video emotion can be expressed with a number of latent emotional events occurring in the

video, and each emotional event reflects in corresponding image features and audio features. For example, there is a user-generated video clip showing that a little girl is scared when she is watching a horror movie. The *scared* is considered to be an emotional event, which simultaneously results in a terrified facial expression in image features, and a scream in audio features. According to this framework, CNN-based deep features and MFCC features respectively represent image features and audio features for video emotion analysis. Sym-cHDP incorporated with a symmetric correspondence mechanism is able to mine such latent emotional events from both deep features and MFCC features within a video. The detailed modeling process will be specified in Section 6.4. Note that deep features and MFCC features must be converted into a number of independent visual words and audio words because of the input of the Sym-cHDP, whose details will be described in Section 6.3.2.

However, the Sym-cHDP is an unsupervised model, a supervised classifier must be joined up after this unsupervised learning process, for predicting a real emotional label for each video in this task.

6.3 Description of Features

6.3.1 CNN-based Deep Features

The latest generation of CNNs, such as AlexNet [36], VGG [37] and GoogleLeNet [38], have achieved remarkable performance for large-scale image classification tasks due to their deep and systematic learning architecture. However, when facing more abstract and complex learning scenario such as image/video emotion recognition, directly using such CNNs cannot perform as good as they did on image classification. But these attempts are still significative, since they provided important clues that the features drawn from high-level (deeper) layers in CNNs potentially contain some sentimental information.

According to the latest study [41], the deep architectures of AlexNet and VGG are more suitable for video emotion recognition than the one of GoogleLeNet. In addition, the features extracted from fully connected layers perform better than the ones extracted from convolutional layers. Therefore, I utilize three different CNN models AlexNet [36], VGG-F [37] and VGG-S [37] in this framework, and respectively extract the deep features from fully connected layers in each of them. Here, VGG-F and VGG-S are two different types of VGG model conditioned on different accuracy/speed trade-offs, where VGG-F with less layers focuses on speed, and VGG-S with more layers focuses on precision. According to the architectures of these three CNN models, the fully connected layers comprise *fc6*

layer and $fc7$ layer, so there are totally 3×2 different combination patterns of deep features evaluated in the experiments.

6.3.2 Extraction of Visual Words and Audio Words

This section specifies how to convert deep features and MFCC features into a number of independent visual words and audio words for Sym-cHDP.

For CNN-based deep features, each deep feature drawn from $fc6$ layer or $fc7$ layer is an m -dimensional vector, which can be expressed with $D_{fj} = (d_{fj1}, d_{fj2}, \dots, d_{fjm})$, where f indicates the index of video, and j indicates the index of image. Here, I assume that each element d maps to an exclusive type of visual word. Hence, the vocabulary size of visual word V_D is equal to m . After *drop-out* processing on fully connected layers, each element d_{fjm} in deep feature vector D_{fj} will have a nonnegative continuous value, which is considered to be a likelihood for extracting the corresponding visual word. Then, I use a multinomial distribution to extract the counts for each visual word:

$$n_{fj} = (n_{fj1}, n_{fj2}, \dots, n_{fjV_D}) \sim \text{Multinomial}(N, P_{D_{fj}}) \quad (6.1)$$

where n_{fj} indicates visual word count vector including V_D types of visual words, and N is the summation of all the counts within n_{fj} , i.e., $N = \sum_{v=1}^{V_D} n_{fjv}$. $P_{D_{fj}}$ is a normalized representation for D_{fj} . Through this method, the visual word set X_{fj} for image j in video f is filled with extracted N visual words. On video level, the visual word set X_f for video f can be obtained with $X_f = \{X_{f1}, X_{f2}, \dots, X_{fJ_f}\}$, where J_f is the number of images in video f .

For audio MFCC features, I assume that each 24-dimensional MFCC descriptor represents an audio word. Then, a simple k -means algorithm is applied to cluster all the MFCC descriptors into V_M clusters, which is treated as the vocabulary size of audio word. Note that an MFCC descriptor is sampled per 5 *ms*, the total audio word counts for each video will depend on the length of the video.

6.4 Symmetric Correspondence HDP

As mentioned in Section 6.2, for a video, the video emotion can be expressed with a number of latent emotional events occurring in the video, and each emotional event reflects in corresponding image features and audio features. To mine such latent emotional events from video multimodal data, this section utilize a Bayesian nonparametric Sym-cHDP [42] model

that incorporates a flexible symmetric correspondence mechanism for modeling multimodal data.

6.4.1 Generative Process

According to the two-layer hierarchy of HDP model, a number of local probability measures $\{G_j\}$ are connected with a shared global probability measure G_0 . These probability measures can be reformulated with corresponding component (topic/emotional event) weights β and π_f through a stick-breaking construction, shown as below:

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\ G_f &= \sum_{k=1}^{\infty} \pi_{fk} \delta_{\phi_k} \end{aligned} \tag{6.2}$$

where I define that the components (topics/emotional events) $\phi = \{\phi_k\}_{k=1}^{\infty}$ are distributed on the base measure H . Each component weight vector consists of an infinite number of corresponding component weights, *i.e.*, $\beta = \{\beta_1, \beta_2, \dots, \beta_k, \dots\}$ and $\pi_f = \{\pi_{f1}, \pi_{f2}, \dots, \pi_{fk}, \dots\}$.

Then, the stick-breaking construction of Sym-cHDP is graphically represented in Fig. 6.2, and its generative process is as follows.

- **Step 1:** Draw the global component weights β from a GEM process [19] $GEM(\gamma)$ with a hyperparameter γ .
- **Step 2:** Conditioned on β , draw each local component weights π_f from a Dirichlet process $DP(\alpha_0, \beta)$ with a hyperparameter α_0 .
- **Step 3:** For each video, draw a multinomial pivot flag generator σ_f from a Dirichlet distribution $Dir(\lambda)$ with a hyperparameter λ .
- **Step 4:** For the i th word of data type¹ (modality) l in video f , draw a pivot flag $s_{fi}^{(l)}$ from a multinomial distribution $Multinomial(\sigma_f)$ conditioned on σ_f .
- **Step 5:** If $s_{fi}^{(l)} = l$, draw an emotional event $z_{fi}^{(l)}$ from π_f . Otherwise if $s_{fi}^{(l)} = h \neq l$, draw an emotional event $z_{fi}^{(l)}$ from a uniform distribution $Uniform(z_{f1}^{(h)}, \dots, z_{fH_f}^{(h)})$.
- **Step 6:** Conditioned on sampled $z_{fi}^{(l)}$, draw a word $x_{fi}^{(l)}$ from a word distribution $f(x_{fi}^{(l)} | \phi_k^{(l)}, k = z_{fi}^{(l)})$.

Similar to the symmetric correspondence in Seq-Sym-cHDP described in Chapter 4, Sym-cHDP also utilizes a multinomial pivot flag generator σ_f to switch the way of generating the

¹Only visual words and audio words are utilized in this work, therefore $l \in \{1, 2\}$.

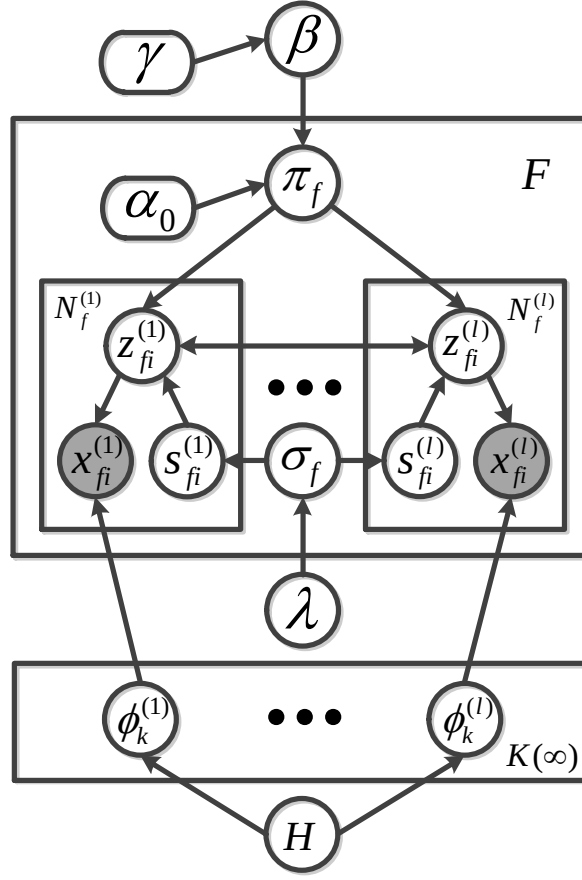


Fig. 6.2 Graphical representation for stick-breaking construction of Sym-cHDP.

emotional event for the current word by sampling a pivot flag. More specifically, for the word $x_{fi}^{(l)}$ within data type l , if the pivot flag generator samples a pivot flag pointing at the same data type, *i.e.*, $s_{fi}^{(l)} = l$, the emotional event $z_{fi}^{(l)}$ for the current word $x_{fi}^{(l)}$ is straightforwardly drawn from π_f . However, the pivot flag generator samples a pivot flag pointing at the other data type h , *i.e.*, $s_{fi}^{(l)} = h \neq l$, the emotional event $z_{fi}^{(l)}$ for the current word $x_{fi}^{(l)}$ is drawn from a uniform distribution that involves all the emotional events assigned to $H_f^{(h)}$ words within data type h . $H_f^{(h)}$ is the number of words that have been already assigned with emotional events at the current step. Hence, $H_f^{(h)} \leq N_f^{(h)}$. Note that all the $\{\phi_k^{(l)}\}_{l=1}^2$ are drawn from a same base measure H , and share the same emotional event space.

With such symmetric correspondence mechanism, Sym-cHDP not only simultaneously handles the generative process of both visual words and audio words within a video, but also flexibly model the dependency between them. This modeling method effectively simulates the expression process of video emotion.

6.4.2 Inference Method

Based on posterior representation sampler [20], I derive an inference method to estimate the latent emotional events and other variables within Sym-cHDP in this section.

Firstly, both of two component weights β and π_f have a finite augmentable representation, which can be sampled with:

$$\begin{aligned}\beta &= (\beta_1, \dots, \beta_K, \beta_u) \sim \text{Dir}(T_1, \dots, T_K, \gamma) \\ \pi_f &= (\pi_{f1}, \dots, \pi_{fK}, \pi_{fu}) \sim \text{Dir}(\tilde{\pi}_{f1}, \dots, \tilde{\pi}_{fK}, \alpha_0 \beta_u)\end{aligned}\quad (6.3)$$

where β_u and π_{fu} are the weight of generating a newborn component within β and π_f , respectively. The variable T_K indicates the table counts in the Chinese restaurant franchise (CRF) representation of Sym-cHDP, whose detailed explanation and sampling approach can be found in Section 3.4.2 of Chapter 3. The estimator $\tilde{\pi}_{fk}$ is derived with: $\tilde{\pi}_{fk} = \alpha_0 \beta_k + \sum_{(l)} C_{f lk}^{(l)}$, where $C_{f lk}^{(l)}$ indicates the counts for the word with data type l in video f possessing the pivot flag with the same data type l when the index of the sampled emotional event is k .

Similar to the posterior representation sampler in Seq-Sym-cHDP, I also derive a full conditional joint likelihood for estimating both emotional event $z_{fi}^{(l)}$ and pivot flag $s_{fi}^{(l)}$ for each word, shown as below:

$$\begin{aligned}P(z_{fi}^{(l)} = k, s_{fi}^{(l)} = l | x_{fi}^{(l)}, \lambda, \dots) &\propto P(s_{fi}^{(l)} = l | \lambda) P(z_{fi}^{(l)} = k | \pi_f) P(x_{fi}^{(l)} | z_{fi}^{(l)} = k, \dots) \\ &= \begin{cases} \frac{C_{fi}^{-fli} + \lambda}{C_{fi}^{-fli} + \sum_{l' \neq l} C_{fi l'} + 2\lambda} \cdot \pi_{fk} \cdot f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) & \text{if } k \text{ is used} \\ \frac{C_{fi}^{-fli} + \lambda}{C_{fi}^{-fli} + \sum_{l' \neq l} C_{fi l'} + 2\lambda} \cdot \pi_{fu} \cdot f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) & \text{if } k \text{ is newborn} \end{cases} \quad (6.4)\end{aligned}$$

$$\begin{aligned}P(z_{fi}^{(l)} = k, s_{fi}^{(l)} = h | x_{fi}^{(l)}, \lambda, \dots) &\propto P(s_{fi}^{(l)} = h | \lambda) P(z_{fi}^{(l)} = k | Z_f^{(h)}) P(x_{fi}^{(l)} | z_{fi}^{(l)} = k, \dots) \\ &= \frac{C_{fh}^{-fli} + \lambda}{C_{fh}^{-fli} + \sum_{h' \neq h} C_{fh h'} + 2\lambda} \cdot \frac{n_{fk}^{(h)}}{N_f^{(h)}} \cdot f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) \quad (6.5)\end{aligned}$$

where C_{fi} indicates the counts for the pivot flag pointing at data type l over all the multimodal data in video f , the superscript $-fli$ for a variable indicates the terms excluding the fli th term. $Z_f^{(h)}$ is a vector that records all the topic assignments for the words of data type h in video f . $n_{fk}^{(h)}$ indicates the counts for emotional event k assigned to words of data type h in video f . Here, $f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ and $f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ represents two different types of conditional word

likelihood functions conditioned on emotional event k . If the emotional event k is previously used, $f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ is formulated with:

$$\begin{aligned} f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)} = v^{(l)}) &= \int f(x_{fi}^{(l)} = v^{(l)} | \phi_k^{(l)}) p(\phi_k^{(l)} | X_{lk}^{-fli}, H) d\phi_k^{(l)} \\ &= \frac{n_{kv^{(l)}}^{-fli} + \tau}{\sum_{v^{(l)}} n_{kv^{(l)}}^{-fli} + V^{(l)} \tau} \end{aligned} \quad (6.6)$$

otherwise if the emotional event k is newborn, $f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ is formulated with:

$$f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) = \int f(x_{fi}^{(l)} = v^{(l)} | \phi_k^{(l)}) p(\phi_k^{(l)} | H) d\phi_k^{(l)} = \frac{1}{V^{(l)}} \quad (6.7)$$

where $v^{(l)}$ is the index of words in the vocabulary of data type l . X_{lk}^{-fli} is a set involving all the words of data type l with emotional event k except for $x_{fi}^{(l)}$. $n_{kv^{(l)}}$ indicates the counts for the word $v^{(l)}$ assigned to emotional event k . $V^{(l)}$ indicates the vocabulary size of data type l . τ is a hyperparameter for a Dirichlet distribution, and there is $H = Dir(\tau)$.

In addition, hyperparameter γ and α_0 can be sampled by setting a gamma prior on them. A Gibbs sampling algorithm is used to implement this inference method, and finally estimates the latent emotional events from visual words and audio words within each video.

6.5 Experiments

6.5.1 Experimental Setup

To evaluate the proposed Bayesian nonparametric multimodal data modeling framework on video emotion recognition, a couple of experiments were conducted for video emotion classification with an *Acted Facial Expressions in the Wild* (AFEW) [43] dataset, which contains 957 video clips extracted from movies. In AFEW dataset, those video clips were categorized with seven basic emotions including *angry*, *happy*, *sad*, *disgust*, *fear*, *surprise* and *neutral*.

For each video clip in AFEW dataset, several key frames involving sentimental information had been extracted in advance. Then, three kinds of CNN models, *i.e.*, AlexNet, VGG-F and VGG-S were utilized to extract corresponding deep features for each key frame of the video. To lower the computation cost for the whole system, the dimension m of *fc6* layer and *fc7* layer in each CNN model was set to 1024, which means the vocabulary size of visual

Table 6.1 Summary of Experimental Dataset

the number of video clips	957
the number of emotion class	7
the number of visual words with <i>fc6</i> or <i>fc7</i>	3666k
the number of audio words	1042k
the vocabulary size of visual words	1024
the vocabulary size of audio words	1000

word V_D was specified to 1024. The total number of extracted visual words N is set to 5120. Other settings on these three CNN models were completely identical to the ones in original works [36, 37]. For audio words extracted from MFCC features, the vocabulary size V_M was set to 1000. The summary of the experimental dataset is shown in Table 6.1.

For the initialization of nonparametric topic models, the hyperparameters γ and α_0 were sampled from a gamma prior $\text{Gamma}(1.0, 1.0)$. The Gibbs sampling system totally ran 1000 iterations so as to let all the variables in the model fully converge.

For the final video emotion classification task, the learned emotional event distribution for each video by nonparametric topic models was input to a simple LIBLINEAR² classifier for final classification. The experiments were conducted and evaluated with a five-fold cross-validation scheme, and the performance was measured with average classification accuracy.

6.5.2 Unimodal Evaluation on CNN-based Deep Features

In this section, I focus on evaluating the performance of video emotion classification with a unimodal HDP by only using different types of CNN-based deep features. For comparison, directly using three different CNN models for video emotion classification were considered to be baselines. The results are shown in Table 6.2.

Evaluation on Deep Features with HDP

HDP as a unimodal version of Sym-cHDP was used for evaluating the video emotion classification performance on different types of deep features. According to the Table 6.2, I respectively set the controlling parameter τ with $\tau = 0.1$, $\tau = 0.5$ and $\tau = 1.0$, to conduct three different sets of experiments. No matter what type of deep feature was utilized, the HDP always performed the best when conditioned on $\tau = 0.5$. For comparing different

²<https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

Table 6.2 Unimodel Results with Different Deep Features

Methods		CNN	HDP($\tau = 0.1$)	HDP($\tau = 0.5$)	HDP($\tau = 1.0$)
AlexNet	fc6	24.68	23.85	24.87	24.13
	fc7		20.53	25.85	24.75
VGG-F	fc6	28.01	26.38	28.84	27.78
	fc7		23.84	26.52	25.14
VGG-S	fc6	31.27	28.11	33.15	31.65
	fc7		26.83	31.35	29.94

Table 6.3 Multimodal Results with Different Methods

CNN Models	AlexNet		VGG-F		VGG-S	
Features	fc6	fc7	fc6	fc7	fc6	fc7
LIBLINEAR	24.11	24.78	28.21	28.54	31.88	30.02
SVM	24.32	25.88	29.65	29.11	32.11	31.88
CI-HDP	25.31	26.16	28.85	27.56	34.10	32.25
Corr-HDP(Visual Words)	25.52	26.53	29.94	28.21	34.34	32.91
Corr-HDP(Audio Words)	24.32	25.52	28.54	27.91	33.56	31.79
Sym-cHDP	27.95	27.89	31.07	29.94	35.22	33.11

types of deep features for each CNN model, the deep features extracted from the *fc6* layer significantly outperformed the ones extracted from the *fc7* layer in the case of using VGG-F and VGG-S, however such phenomenon was not so clear in the case of using AlexNet.

Comparison between CNN methods and HDP methods

I also respectively conducted three comparative experiments by directly using three corresponding CNN models as baselines. Compared with baseline results, the optimized HDP method outperformed CNN method by 4.47%, 2.96% and 6.01%, when using AlexNet, VGG-F and VGG-S, respectively. This demonstrates that HDP method is able to improve the performance of video emotion recognition via mining latent emotional events from deep features within videos. On the other hand, VGG-S achieved the best performance on video emotion classification among all the CNN models, which implies that the VGG-S architecture is more suitable for the emotion recognition task.

6.5.3 Multimodal Evaluation on Modeling Methods

In this section, comparative experiments based on video multimodal data (deep features and MFCC features) were conducted for assessing the proposed method and other baselines on video emotion classification. Two styles of baseline sets were adopted for evaluation: the baseline models in the first set were general classifiers including LIBLINEAR and SVM, where the deep features and MFCC features were simply fused without being further mined, the baseline models in the second set were other nonparametric topic models incorporated in the same framework including conditionally independent HDP (CI-HDP) deemed as two-layer version of Seq-CI-HDP, and correspondence HDP (Corr-HDP) deemed as two-layer version of Seq-cHDP. Similar to Seq-cHDP, Corr-HDP also had two different scenarios with respectively predefining the pivot to visual words and audio words in the experiments. In initialization, I set $\tau = 0.5$ to all the nonparametric topic models, and set $\lambda = 1.0$ to Sym-cHDP. The experimental results are shown in Table 6.3.

Compared with both LIBLINEAR and SVM, the proposed method with Sym-cHDP significantly outperformed those general classifiers for every scenario, which demonstrates that Sym-cHDP is able to improve the performance of video emotion recognition via mining latent emotional events from both deep features and MFCC features within videos. On the other hand, Sym-cHDP outperformed other nonparametric topic models CI-HDP and Corr-HDP for every scenario as well, which demonstrates that incorporating such flexible symmetric correspondence mechanism in a nonparametric multimodal topic model can improve the accuracy of video emotion recognition. The best performance among all the scenarios was achieved when the proposed framework worked with Sym-cHDP conditioned on using the deep features extracted from the *fc6* layer of VGG-S.

Compared with the results shown in Table 6.2, I also found that Sym-cHDP performing on multimodal data significantly outperformed HDP performing on unimodal data (deep features) for every scenario, which demonstrates that the MFCC features are very complementary to the deep features for such video emotion recognition task, since they express the emotional information from different angles within a video.

6.6 Summary

This chapter proposed a Bayesian nonparametric multimodal data modeling framework to deeply learn the video emotion from CNN-based deep features and MFCC features extracted from videos. In this framework, a symmetric correspondence hierarchical Dirichlet processes

(Sym-cHDP) [42] is utilized to model the multimodal data, and mine the latent emotional events (topics) from both deep features and MFCC features.

According to this work, there are two major contributions: (1) video emotion recognition by using CNN-based deep features is a very state-of-the-art work, where I conducted a number of comprehensive experiments to explore the most appropriate CNN architecture and deep feature for learning emotion from a video, (2) the nonparametric multimodal topic modeling method, for the first time, was utilized for mining the latent emotional events from video multimodal data, which offers a new solution for video emotion analysis.

In future works, similar to the CNN-based deep features, deeply learned audio features extracted from other neural networks are also promising to be utilized in this video emotion recognition framework. Furthermore, a supervised nonparametric topic model is expected to be incorporated into this framework to improve the effectiveness of video emotion classification.

Chapter 7

Conclusion

This paper focused on video data analysis with Bayesian nonparametric topic modeling methods. According to the characteristic of video data and application scenario, I developed a number of Bayesian nonparametric topic models to tackle the corresponding problems, and demonstrated their effectiveness via comprehensive experiments. I respectively conclude all the works in the following.

Chapter 3 presented a sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) model for video semantic analysis, which solved two major problems: (1) a sequence of video frames within each video are modeled via a time-dependency mechanism based on a Markovian assumption, (2) video multimodal data are modeled via a corresponding mechanism. To infer the latent topics from observed video data with Seq-cHDP, I also derived an inference method based on posterior representation samplers, and implemented it with a cascaded Gibbs sampling algorithm. In experimentation, I performed video semantic analysis based on extracted SIFT visual words and speech transcript words, and finally evaluated the Seq-cHDP model with topic trend estimation and genre classification.

Chapter 4 firstly discussed two simple multimodal data modeling methods for video data that respectively occurred in Seq-cHDP and sequential conditionally independent hierarchical Dirichlet processes (Seq-CI-HDP). To improve the weaknesses of these two multimodal data modeling mechanisms, I presented a sequential symmetric correspondence hierarchical Dirichlet processes (Seq-Sym-cHDP) model for the same purpose, which incorporated a symmetric correspondence mechanism that can flexibly model the dependency among multimodal data with pivot assignments. The experiments demonstrated that the Seq-Sym-cHDP outperformed Seq-cHDP and Seq-CI-HDP with both qualitative evaluation and quantitative evaluation.

Chapter 5 presented a supervised learning version of Seq-Sym-cHDP named supervised sequential symmetric correspondence hierarchical Dirichlet processes (Sup-SSC-HDP) for multi-class video classification based on video semantic analysis. Sup-SSC-HDP utilized a generalized linear model (GLM) to model the relationship between empirical topic frequencies over multimodal data and video class. For inference, a softmax regression algorithm is applied to optimize the estimate of their regression coefficients. The experiments demonstrated that such supervised learning method can give a uplift on the performance of video classification compared with corresponding unsupervised model.

Chapter 6 presented a Bayesian nonparametric multimodal data modeling framework for learning the human emotion from video data. This framework incorporated two core models: (1) a convolutional neural network (CNN) model for deeply learning the images extracted from each video, (2) a two-layer symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) model for mining latent emotional events (topics) from CNN-based deep features and MFCC audio features. In experimentation, I conducted a number of comprehensive experiments to explore the most appropriate CNN architecture and deep feature for learning emotion from a video, meanwhile also demonstrated the effectiveness of this framework on video emotion classification compared with other baseline methods.

In future works, I will consider to improve the current time-dependency mechanism incorporated in proposed sequential nonparametric topic models with a more sophisticated and flexible approach, for example, hidden Markov models (HMMs), which might increase the modeling accuracy without performing parameter tuning. Furthermore, for the supervised learning model Sup-SSC-HDP, a more effective supervised learning method such as maximum entropy discrimination [29] is expected to be utilized in Sup-SSC-HDP to further improve the performance of video classification. In addition, the proposed Bayesian nonparametric topic models are also very promising to be extended to other application scenarios other than video semantic analysis and video emotion recognition.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Koji Eguchi for the continuous support of my Ph.D study, for his patience, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides, I would like to thank the rest of my thesis committee: Prof. Takenao Ohkawa, Prof. Kuniaki Uehara and Prof. Hisashi Tamaki, for their encouragement, insightful comments, and valuable questions.

My sincere thanks also go to other fellow labmates in CS-25 laboratory of Kobe University, for their assistance, cooperation and of course friendship.

Last but not the least, I would like to thank my family: my parents Zhigang Xue and Huiling Li, for giving birth to me at the first place and supporting me spiritually throughout my life.

References

- [1] P. Orbanz and Y.W. Teh, “Bayesian nonparametric models,” in Encyclopedia of Machine Learning, pp.81–89, Springer, 2011.
- [2] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet allocation,” the Journal of machine Learning research, vol.3, pp.993–1022, 2003.
- [3] Y.W. Teh, K. Kurihara, and M. Welling, “Collapsed variational inference for hdp,” Advances in neural information processing systems, pp.1481–1488, 2008.
- [4] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, “Hierarchical Dirichlet processes,” Journal of the american statistical association, vol.101, no.476, 2006.
- [5] T.L. Griffiths and M. Steyvers, “Finding scientific topics,” Proceedings of the National academy of Sciences, vol.101, no.suppl 1, pp.5228–5235, 2004.
- [6] Y.W. Teh, “Dirichlet process,” in Encyclopedia of machine learning, pp.280–287, Springer, 2010.
- [7] X. Wang, X. Ma, and E. Grimson, “Unsupervised activity perception by hierarchical Bayesian models,” IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007, pp.1–8, IEEE, 2007.
- [8] T. Hospedales, S. Gong, and T. Xiang, “A Markov clustering topic model for mining behaviour in video,” IEEE 12th International Conference on Computer Vision, 2009, pp.1165–1172, IEEE, 2009.
- [9] D. Kuettel, M.D. Breitenstein, L. Van Gool, and V. Ferrari, “What’s going on? discovering spatio-temporal dependencies in dynamic scenes,” IEEE Conference on Computer Vision and Pattern Recognition, 2010. CVPR 2010., pp.1951–1958, IEEE, 2010.
- [10] P. Das, C. Xu, R.F. Doell, and J.J. Corso, “A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching,” IEEE

- Conference on Computer Vision and Pattern Recognition (CVPR), 2013. CVPR 2013, pp.2634–2641, IEEE, 2013.
- [11] X. Yang and K. Eguchi, “Multimedia topic models considering burstiness of local features,” *IEICE TRANSACTIONS on Information and Systems*, vol.97, no.4, pp.714–720, 2014.
- [12] L. Ren, D.B. Dunson, and L. Carin, “The dynamic hierarchical dirichlet process,” *Proceedings of the 25th international conference on Machine learning*, pp.824–831, ACM, 2008.
- [13] J. Zhang, Y. Song, C. Zhang, and S. Liu, “Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora,” *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.1079–1088, ACM, 2010.
- [14] D.M. Blei and M.I. Jordan, “Modeling annotated data,” *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp.127–134, ACM, 2003.
- [15] K. Fukumasu, K. Eguchi, and E.P. Xing, “Symmetric correspondence topic models for multilingual text analysis,” *Advances in Neural Information Processing Systems*, pp.1286–1294, 2012.
- [16] J.D. Mcauliffe and D.M. Blei, “Supervised topic models,” *Advances in neural information processing systems*, pp.121–128, 2008.
- [17] A.M. Dai and A.J. Storkey, “The supervised hierarchical dirichlet process,” *IEEE transactions on pattern analysis and machine intelligence*, vol.37, no.2, pp.243–255, 2015.
- [18] J. Sethuraman, “A constructive definition of Dirichlet priors,” tech. rep., DTIC Document, 1991.
- [19] J. Pitman, “Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition,” *Combinatorics, Probability and Computing*, vol.11, no.5, pp.501–514, 2002.
- [20] Y.W. Teh and M.I. Jordan, “Hierarchical Bayesian nonparametric models with applications,” *Bayesian nonparametrics*, vol.1, 2010.

- [21] I. Porteous, Networks of Mixture Blocks for Non Parametric Bayesian Models with Applications DISSERTATION, Ph.D. thesis, University of California, Irvine, 2010.
- [22] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G.J. Jones, "Overview of MediaEval 2011 rich speech retrieval task and genre tagging task," CEUR Workshop Proceedings, 2011.
- [23] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol.60, no.2, pp.91–110, 2004.
- [24] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005., pp.524–531, IEEE, 2005.
- [25] J.P. Callan, W.B. Croft, and S.M. Harding, "The INQUERY retrieval system," Database and expert systems applications, pp.78–83, Springer, 1992.
- [26] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," ACM transactions on intelligent systems and technology (TIST), vol.2, no.3, p.27, 2011.
- [27] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," Proceedings of the National Academy of Sciences, vol.101, no.suppl 1, pp.5220–5227, 2004.
- [28] L.v.d. Maaten and G. Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research, vol.9, no.Nov, pp.2579–2605, 2008.
- [29] J. Zhu, A. Ahmed, and E.P. Xing, "Medlda: maximum margin supervised topic models," Journal of Machine Learning Research, vol.13, no.Aug, pp.2237–2278, 2012.
- [30] S.C. Marsella and J. Gratch, "Ema: A process model of appraisal dynamics," Cognitive Systems Research, vol.10, no.1, pp.70–90, 2009.
- [31] H.B. Kang, "Affective content detection using hmms," Proceedings of the eleventh ACM international conference on Multimedia, pp.259–262, ACM, 2003.
- [32] M. Xu, C. Xu, X. He, J.S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," Signal Processing, vol.93, no.8, pp.2140–2150, 2013.

- [33] R.M.A. Teixeira, T. Yamasaki, and K. Aizawa, “Determination of emotional content of video clips by low-level audiovisual features,” *Multimedia Tools and Applications*, vol.61, no.1, pp.21–49, 2012.
- [34] Y.G. Jiang, B. Xu, and X. Xue, “Predicting emotions in user-generated videos,” *AAAI*, pp.73–79, 2014.
- [35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol.86, no.11, pp.2278–2324, 1998.
- [36] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, pp.1097–1105, 2012.
- [37] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” *Proceedings of the British Machine Vision Conference*, 2014.
- [38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–9, 2015.
- [39] T. Chen, D. Borth, T. Darrell, and S.F. Chang, “Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks,” *CoRR*, 2014.
- [40] Q. You, J. Luo, H. Jin, and J. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp.381–388, AAAI Press, 2015.
- [41] B. Xu, Y. Fu, Y.G. Jiang, B. Li, and L. Sigal, “Video emotion recognition with transferred deep feature encodings,” *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp.15–22, ACM, 2016.
- [42] K. Eguchi, K. Fukumasu, T. Omoto, and E.P. Xing, “Symmetric correspondence topic models for multimodal data analysis,” *ArXiv*.
- [43] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, “Acted facial expressions in the wild database,” *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, vol.2, 2011.

Publications

Journal Papers

1. Jianfei Xue and Koji Eguchi, Video Data Modeling using Sequential Correspondence Hierarchical Dirichlet Processes, *IEICE Transactions on Information and Systems*, Vol. E100-D, No. 1, pp.33-41, 2017.
2. Jianfei Xue and Koji Eguchi, Sequential Bayesian Nonparametric Multimodal Topic Models for Video Data Analysis, *IEICE Transactions on Information and Systems* (To Appear).

Conference Papers

1. Jianfei Xue and Koji Eguchi, Sequential Correspondence Hierarchical Dirichlet Processes for Video Data Analysis, *Proceedings of the 6th ACM International Conference on Multimedia Retrieval (ICMR 2016)*, pp.229-233, 2016.
2. Jianfei Xue and Koji Eguchi, Sequential Symmetric Correspondence Hierarchical Dirichlet Processes for Video Data Analysis, *Advances in Neural Information Processing Systems (NIPS 2016), Workshop on Practical Bayesian Nonparametrics*, 2016.
3. Jianfei Xue, Zhaojie Luo, Koji Eguchi, Tetsuya Takiguchi and Tsukasa Omoto, A Bayesian Nonparametric Multimodal Data Modeling Framework for Video Emotion Recognition, *2017 IEEE International Conference on Multimedia and Expo (ICME 2017)*, pp. 601-606, 2017.
4. Jianfei Xue and Koji Eguchi, Supervised Nonparametric Multimodal Topic Modeling Methods for Multi-class Video Classification, *The 8th ACM International Conference on Multimedia Retrieval (ICMR 2018)* (Under Review).

Others

1. Jianfei Xue and Koji Eguchi, Sequential Correspondence Hierarchical Dirichlet Processes for Video Data Analysis, *2016 Kobe University Core-Team Workshop on Cyber-Physical System for Smarter World*, Poster Presentation, 2016.
2. Jianfei Xue and Koji Eguchi, Sequential Symmetric Correspondence Hierarchical Dirichlet Processes for Video Data Analysis, *2016 Information-Based Induction Sciences Workshop, (IBIS Workshop 2016)*, Poster Presentation, 2016.
3. Jianfei Xue and Koji Eguchi, Supervised Nonparametric Multimodal Topic Modeling Methods for Multi-class Video Classification, *2017 Information-Based Induction Sciences Workshop, (IBIS Workshop 2017)*, Poster Presentation, 2017.

Doctor Thesis, Kobe University

“Bayesian Nonparametric Topic Models for Video Data Analysis ”, 94 pages

Submitted on January, 24th, 2018

The date of publication is printed in cover of repository version published in

Kobe University Repository Kernel.

© XUE JIANFEI

All Right Reserved, 2018