PAPER Special Section on Enriched Multimedia —New Technology Trends in Creation, Utilization and Protection of Multimedia Information—

Video Data Modeling Using Sequential Correspondence Hierarchical Dirichlet Processes

Jianfei XUE[†a)], ***Nonmember and*** **Koji EGUCHI**[†b)], ***Member***

SUMMARY Video data mining based on topic models as an emerging technique recently has become a very popular research topic. In this paper, we present a novel topic model named sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) to learn the hidden structure within video data. The Seq-cHDP model can be deemed as an extended hierarchical Dirichlet processes (HDP) model containing two important features: one is the time-dependency mechanism that connects neighboring video frames on the basis of a time dependent Markovian assumption, and the other is the correspondence mechanism that provides a solution for dealing with the multimodal data such as the mixture of visual words and speech words extracted from video files. A cascaded Gibbs sampling method is applied for implementing the inference task of Seq-cHDP. We present a comprehensive evaluation for Seq-cHDP through experimentation and finally demonstrate that Seq-cHDP outperforms other baseline models.

key words: *bayesian nonparametric methods, multimedia machine learning, hierarchical Dirichlet processes, topic models*

1. Introduction

With the advances in information and communication technology, nowadays multimedia services have been playing an ever more important role on the Internet. Meanwhile, these contents and services are becoming increasingly varied. For instance, now we can share any video clips through a video website such as YouTube while attaching descriptions to let viewers easily find clips that interest them. However, an anonymous video file without any descriptions can also be uploaded to the video website without being sorted into a specific genre. As the matter of fact, these anonymous videos need to be analyzed properly, since they may contain very useful and important information. To analyze a video from the perspective of content, the core issue is how to learn and summarize the latent semantic information from multimedia data within the video, which is also the main purpose of this paper.

Image and video data mining combined with topic modeling methods such as Latent Dirichlet Allocation (LDA) [1] and hierarchical Dirichlet processes (HDP) [2] has recently been attracting more and more focus in this field. Generally, topic models are based on the 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. Therefore, some features or information extracted from an image or a video file can also be considered as visual words corresponding to the topic models. However, straightforwardly applying LDA or HDP to deal with video data is not a proper solution, since video data has a more complicated structure (multimodal data) than text data.

A video file can be deemed as a type of multimodal data, which generally contains image information and speech information. To be more specific, the image information stands for a sequence of video frames, while the speech information stands for speech transcript. To cope with this kind of video data, in this paper, we propose a sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) model, which is a kind of modified HDP model involving two important aspects. One is the time-dependency mechanism that shows the time relativity between neighboring video frames, and the other is the data correspondence mechanism that provides two corresponding generative processes for the multimodal data within the model. Then, a cascaded Gibbs sampler is employed for inferring the Seq-cHDP model. In the experimental phase, genre classification is performed to evaluate the Seq-cHDP model. We demonstrate that our model outperforms the other baselines by showing the experimental results of the topic trend estimation and classification accuracy.

2. Related Work

As one of the key topic models, the HDP model was firstly proposed by Teh et al. [2]. They provided a new theory that models multiple correlated data corpora as multiple infinite Dirichlet processes (DP) [3], and connects them via sharing mixing components among all corpora. On the basis of this theory, various of applications have been widely developed for data mining. In this paper, we also employ HDP as the basic model to extend it for video data analysis.

Many recent works have addressed image or video data mining by using topic models [4]–[10]. In previous work on video data mining, Souvannavong et al. developed an efficient framework on the basis of probabilistic latent semantic analysis (PLSA) for video shot indexing and retrieval [7]. However, only visual information is considered in their work. Wang et al. presented a novel unsupervised learning method on the basis of both the LDA and HDP models to detect activities and interactions that occur in videos [8]. However, the drawback of their work is

Manuscript received April 5, 2016.
Manuscript revised August 1, 2016.
Manuscript publicized October 7, 2016.
[†]The authors are with the Graduate School of System Informatics, Kobe University, Kobe-shi, 657–8501 Japan.
a) E-mail: silxue@cs25.scitec.kobe-u.ac.jp
b) E-mail: eguchi@port.kobe-u.ac.jp
DOI: 10.1587/transinf.2016MUP0007

that they ignored the time-dependency feature within video data. Hospedales et al. proposed a novel Markov clustering topic model (MCTM) for unsupervised learning of scene characteristics, dynamically screening and identifying irregular spatiotemporal patterns in video [9]. However, their model is extended from LDA, which lacks flexibility compared with HDP, since the initial number of the topics set in LDA affects the final performance. Kuettel et al. presented a cascaded topic model called dependent Dirichlet processes and hidden Markov models (DDP-HMM) to jointly learn spatio-temporal dependencies of moving agents in complex dynamic scenes [10]. However, it is difficult to find a way to carry out multimodal data modelling based on the structure of DDP-HMM. Yang et al. exploited a new topic model called correspondence Dirichlet compound multinomial LDA (Corr-DCMLDA), which incorporates Dirichlet compound multinomial LDA (DCMLDA) [11] into correspondence LDA (CorrLDA) [4] to tackle the burstiness problem of the local features for video data mining. However, they did not consider the time-dependency issue either. Moreover, the concerns about the Corr-DCMLDA model lacking flexibility still exist.

3. Model

The features of video data are distinguished from those of general text data. However, we can borrow the idea of processing the text data by applying the HDP model. According to the features of video data, a video file consists of amounts of sequential video frames, which can be considered as a sequence of images. For these frames, not only can visual information be observed but also speech information can be extracted by speech recognition techniques. From the perspective of topic models, we can consider that a video file consists of visual words and speech words, which are filled in every frame of a video file.

In this section, we first develop a unimodal sequential HDP (Seq-HDP) model to deal with the time-dependency issue about neighboring video frames. Next, a stick-breaking construction for Seq-HDP is described. Then, the Seq-cHDP model incorporated with the idea of correspondence method is proposed for multimodal data. Finally, we present a posterior representation scheme for inferring the Seq-cHDP.

3.1 Sequential HDP

Borrowing the idea from [12] utilized for modeling the time-varying activities, here we present a simple three-layer HDP model named Seq-HDP to show the time dependencies among neighboring frames within each video file. Figure 1 shows a graphical representation of Seq-HDP. We define base measure G placed in the top of the model structure as an overall measure, which is simultaneously shared by all the HDPs. Similarly, on the second layer, G_0^f denotes the global measure of the f-th video file, and on the third layer, G_j^f denotes the local measure in the j-th frame of the f-th video file.

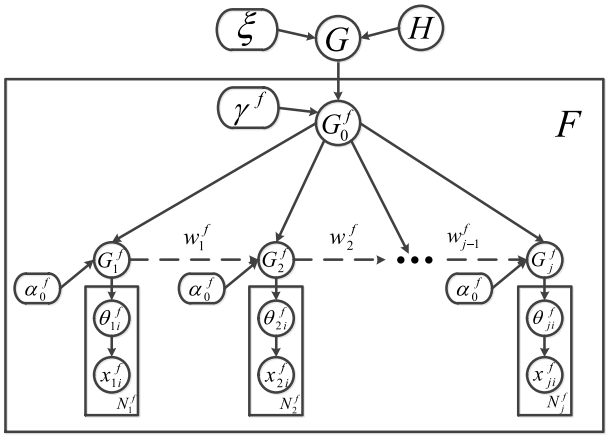


Fig. 1 Graphical representation for Seq-HDP.

However, according to the features of a video file, video frames belonging to the same video file are sequential and time related. To deal with this issue, we set up a dependent relationship between two adjacent local measures G_{j-1}^f and G_j^f in the Seq-HDP model, and we consider this dependency is on the basis of a Markovian assumption. Therefore, as shown in the graphical representation of Seq-HDP, we use transfer weight w_j^f for balancing the generative process of G_j^f, which is specified in the following paragraphs.

Then, the generative process of Seq-HDP is described as follows.

1. The overall measure G recording all the components within the whole model is drawn from $G \sim DP(\xi, H)$. Here, H is a base measure, and ξ is a concentration parameter. DP stands for a Dirichlet process [13].

2. Each video file contains a global measure G_0^f, which is drawn from $G_0^f \sim DP(\gamma^f, G)$. Here, f is the index of the video file, and γ^f is a concentration parameter for generating the G_0^f.

3. Then, we use G_j^f to represent each local measure, and the local measures are time dependent. Therefore, the generation of G_j^f can be affected by both the upper global measure G_0^f and previous local measure G_{j-1}^f.

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

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

4. Finally, the data sample x_{ji}^f can be extracted after drawing the parameters of the component densities θ_{ji}^f, it can be obtained as below:

$$\theta_{ji}^f \sim G_j^f, \quad x_{ji}^f \sim F(x|\theta_{ji}^f) \tag{2}$$

where i denotes the index of the data sample, and $F(x|\theta_{ji}^f)$ is a distribution with parameter θ_{ji}^f.

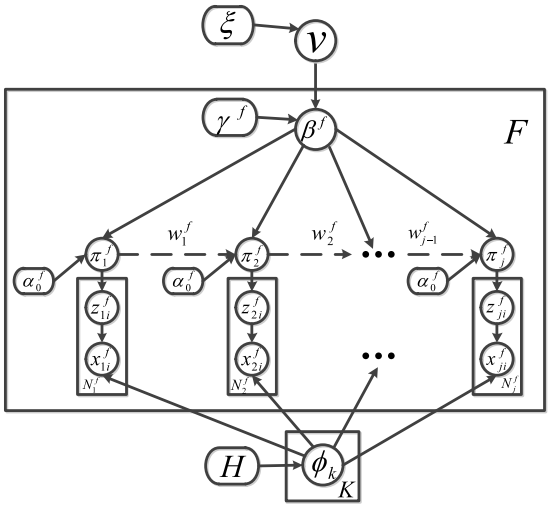


Fig. 2 Graphical representation for stick-breaking construction of Seq-HDP.

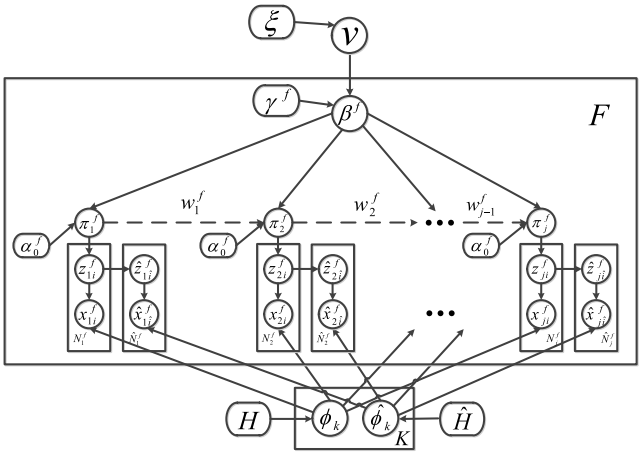


Fig. 3 Graphical representation for stick-breaking construction of Seq-cHDP.

3.2 The Stick-Breaking Construction for Seq-HDP

To better understand the processes of generating samples and allocating components through the whole Seq-HDP, we introduce the stick-breaking construction [2] to provide another description of the model from another perspective.

Figure 2 depicts the stick-breaking construction of Seq-HDP. According to Sethuraman's theory [14] about stick-breaking construction for HDP, the overall measure G can be constructed as below:

$$G = \sum_{k=1}^{\infty} \nu_k \delta_{\phi_k}, \quad \boldsymbol{\nu} \sim GEM(\xi) \tag{3}$$

where we use infinite vector $\boldsymbol{\nu}$ to collect overall component weight ν_k by $\boldsymbol{\nu} = \{\nu_1, \nu_2, \ldots\}$. Moreover, $\boldsymbol{\nu}$ can be obtained through a GEM distribution [2], which is specified by such a process: $\hat{\nu}_k \sim Beta(1, \xi)$, $\nu_k = \hat{\nu}_k \Pi_{i=1}^{k-1}(1 - \hat{\nu}_i)$. δ_{ϕ_k} is a probability measure on component ϕ_k.

Similarly, the global measure G_0^f is formed with:

$$G_0^f = \sum_{k=1}^{\infty} \beta_k^f \delta_{\phi_k}, \quad \boldsymbol{\beta^f} \sim DP(\gamma^f, \boldsymbol{\nu}) \tag{4}$$

where β_k^f is the global component weight corresponding to G_0^f and $\boldsymbol{\beta^f} = \{\beta_1^f, \beta_2^f, \ldots\}$.

Then, on the basis of Eq. (1), we give the form of local measure G_j^f:

$$G_j^f = \sum_{k=1}^{\infty} \pi_{jk}^f \delta_{\phi_k}, \quad \boldsymbol{\pi_j^f} \sim DP(\alpha_0^f, \boldsymbol{\pi'^f_j}) \tag{5}$$

where π_{jk}^f is the local component weight corresponding to G_j^f and $\boldsymbol{\pi_j^f} = \{\pi_{j1}^f, \pi_{j2}^f, \ldots\}$, because of the effect of dependency, $\boldsymbol{\pi'^f_j}$ can be expressed as $\boldsymbol{\pi'^f_j} = w_{j-1}^f \boldsymbol{\pi_{j-1}^f} + (1 - w_{j-1}^f)\boldsymbol{\beta^f}$.

Through this approach, we can obtain the stick-breaking construction for Seq-HDP. Note that z_{ji}^f drawn from π_{jk}^f is the index of the component, which determines the generation of sample x_{ji}^f.

3.3 Correspondence Method

As we mentioned in the beginning of this section, sometimes, the video file involves some text information such as speech transcript, which can be considered as additional featured information depending on the video data. To cope with this complex situation, we incorporate the LDA-based correspondence method proposed by Blei and Jordan [4] into our task and name our new model the sequential correspondence hierarchical Dirichlet processes (Seq-cHDP).

Figure 3 depicts the stick-breaking construction of Seq-cHDP. According to the perspective of topic model, 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 $\boldsymbol{\pi_j^f}$. The rest of the generation process of Seq-cHDP is given below.

1. For each frame j, draw a topic z_{ji}^f for a visual word from $\boldsymbol{\pi_j^f}$. Then in accordance with the sampled topic z_{ji}^f, a specific visual word x_{ji}^f can be drawn from $f(x_{ji}^f | \phi_k, k = z_{ji}^f)$.

2. In the same frame, draw a topic $\hat{z}_{ji}^f$ for a speech word from $Uniform(z_{j1}^f, \ldots, z_{jN_j^f}^f)$. Then in accordance with the sampled topic $\hat{z}_{ji}^f$, a specific speech word $\hat{x}_{ji}^f$ can be drawn from $f(\hat{x}_{ji}^f | \hat{\phi}_k, k = \hat{z}_{ji}^f)$.

Note that the process of drawing a topic of a speech word is uncorrelated with topic weight $\boldsymbol{\pi_j^f}$, and that it only depends on the sampled topic counts of visual words in the same frame. N_j^f and $\hat{N}_j^f$ denotes the number of visual words

and speech words, respectively. $f(\cdot|\cdot)$ denotes a conditional distribution. In fact, there are two different overall topic density parameters ϕ_k and $\hat{\phi}_k$ that are respectively drawn from two different base measure H and $\hat{H}$ in this model and respectively belong to visual words and speech words. However, they share the same K topics.

3.4 Posterior Representation for Seq-cHDP

In this section, we employ a cascaded Gibbs sampling method for Seq-cHDP inference. This inference scheme is based on a posterior representation sampler of the original HDP [15].

3.4.1 A Chinese Restaurant Metaphor

In the original HDP, the Chinese restaurant franchise (CRF) was developed for understanding the HDP from another perspective. According to the scenario of CRF, there are several Chinese restaurants serving customers dishes from the same menu. Here, a customer indicates a single word, a dish indicates a latent topic, and the menu stands for the collection of topics. We assume that each Chinese restaurant has 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 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. She can sit at the same table if she is willing to eat the same dish as the previous customer. Otherwise, she can sit at another table and order any dish she wants. Note that the dish can also be a new dish if she is not interested in the dishes written on the global menu. Following these steps, the whole process of the CRF can be deemed as a sort of generative process of HDP.

However, in the situation of three-layer Seq-cHDP, the generative process is more complicated. According to the theory in paper [12], metatables (tables of the tables) are introduced for constructing the Dirichlet process of the overall measure G. In the Seq-cHDP model, it has a similar structure to CRF. Therefore, based on the method mentioned in paper [12], the extracted counts of metatables, tables and customers can be utilized for estimating the component weighting parameter $\boldsymbol{\nu}$, $\boldsymbol{\beta}^f$ and $\boldsymbol{\pi}^f_j$. Figure 4 shows the brief generation mechanism of stick-breaking construction for Seq-cHDP associated with the extracted counts of CRF. M^f_k denotes the number of the metatables in the area (file) f serving dish k which is drawn from the overall menu $\boldsymbol{\nu}$. $T^f_{(j+1)k}$ denotes the number of the tables in the restaurant $j+1$ of area (file) f serving dish k. However, this dish can be selected from the previous local menu $\boldsymbol{\pi}^f_j$ or the global menu $\boldsymbol{\beta}^f$ due to time dependency. Therefore, we divide $T^f_{(j+1)k}$ into two types by $T^f_{(j+1)k} = T^{j\to j+1}_{fk} + T^{0\to j+1}_{fk}$, where $T^{j\to j+1}_{fk}$ represents the counts of tables whose dishes are drawn from $\boldsymbol{\pi}^f_j$, and $T^{0\to j+1}_{fk}$ represents the counts of tables whose dishes are drawn from $\boldsymbol{\beta}^f$. Then, n^f_{jk} denotes the number of the customers eating dish k in the restaurant j of area (file) f, and the dishes they ordered are from the local menu $\boldsymbol{\pi}^f_j$. Note that all the topic index parameters k on counts are marginalized with dots in Fig. 4.

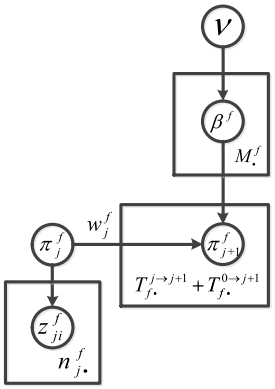


Fig. 4 Generation mechanism of Seq-cHDP on stick-breaking construction associated with the extracted counts of CRF.

3.4.2 Variables Sampling

In this section, we will describe the way to sample the variables mentioned in the stick-breaking construction for Seq-cHDP.

According to the scheme of the posterior representation sampler [15], the overall measure G can be formed by:

$$G = \sum_{k=1}^{K} \nu_k \delta_{\phi_k} + \nu_u G_u, \quad G_u \sim DP(\xi, H) \tag{6}$$

and the vector $\boldsymbol{\nu}$ can be estimated by:

$$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K, \nu_u) \sim Dir(M^{\cdot}_1, \ldots, M^{\cdot}_K, \xi) \tag{7}$$

where Dir denotes Dirichlet distribution. As described in Sect. 3.4.1, $M^{\cdot}_k$ is the marginal form of M^f_k, hence $M^{\cdot}_k = \sum_f M^f_k$. In this procedure, the original infinite vector $\boldsymbol{\nu}$ has an augmentable form filled with finite K components and a promising component u.

Similarly, the global measure G^f_0 can be formed by:

$$G^f_0 = \sum_{k=1}^{K} \beta^f_k \delta_{\phi_k} + \beta^f_u G_u, \quad G_u \sim DP(\xi, H) \tag{8}$$

and the vector $\boldsymbol{\beta}^f$ can be estimated by:

$$\boldsymbol{\beta}^f = (\beta^f_1, \ldots, \beta^f_K, \beta^f_u) \sim Dir(\tilde{\beta}^f_1, \ldots, \tilde{\beta}^f_K, \tilde{\beta}^f_u) \tag{9}$$

where $\tilde{\beta}^f_k = \gamma^f \nu_k + T^{0\to\cdot}_{fk}$ and $\tilde{\beta}^f_u = \gamma^f \nu_u$. According to the CRF of Seq-cHDP, only some of the tables $T^{0\to\cdot}_{fk}$ assigned with dish k among all the restaurants are associated with β^f_k.

Here, $T_{fk}^{0\to\cdot} = \sum_j T_{fk}^{0\to j}$.

Then, the local measure G_j^f is expressed as follows:

$$G_j^f = \sum_{k=1}^{K} \pi_{jk}^f \delta_{\phi_k} + \pi_{ju}^f G_u, \quad G_u \sim DP(\xi, H) \tag{10}$$

and the vector $\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{f}}$ can be estimated by:

$$\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{f}} = (\pi_{j1}^f, \ldots, \pi_{jK}^f, \pi_{ju}^f) \sim Dir(\tilde{\pi}_{j1}^f, \ldots, \tilde{\pi}_{jK}^f, \tilde{\pi}_{ju}^f) \tag{11}$$

as can be seen in Fig. 4, the parameter $\boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{f}}$ has two generating paths that respectively generate $n_{j\cdot}^f$ and $T_{f\cdot}^{j\to j+1}$. Hence, the estimator $\tilde{\pi}_{jk}^f$ and $\tilde{\pi}_{ju}^f$ can be expressed as follows:

$$\tilde{\pi}_{jk}^f = \alpha_0^f w_{j-1}^f \pi_{(j-1)k}^f + \alpha_0^f (1 - w_{j-1}^f)\beta_k^f + n_{jk}^f + T_{fk}^{j\to j+1} \tag{12}$$

$$\tilde{\pi}_{ju}^f = \alpha_0^f w_{j-1}^f \pi_{(j-1)u}^f + \alpha_0^f (1 - w_{j-1}^f)\beta_u^f \tag{13}$$

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

In addition, the concentration parameters ξ, γ^f and α_0^f corresponding to the different layers can be sampled by setting a gamma prior on them:

$$\xi \sim Ga(a_\xi, b_\xi);\ \gamma^f \sim Ga(a_\gamma, b_\gamma);\ \alpha_0^f \sim Ga(a_{\alpha_0}, b_{\alpha_0}) \tag{14}$$

where all the shaping parameters $a.$ and scaling parameters $b.$ are given before sampling. This method is also specified in [12].

In the following, we will introduce the sampling approach of table counts T_{jk}^f and metatable counts M_k^f, which are very necessary parameters for sampling component weights as mentioned above.

Referring to paper [16], Porteous proposes a Bernoulli trial method for replacing the Chinese restaurant process (CRP) to estimate the table counts. On the basis of his approach, we give the estimation of T_{jk}^f:

$$p(T_{jkr}^f = 1) = \frac{\alpha_0^f [w_{j-1}^f \pi_{(j-1)k}^f + (1 - w_{j-1}^f)\beta_k^f]}{\alpha_0^f [w_{j-1}^f \pi_{(j-1)k}^f + (1 - w_{j-1}^f)\beta_k^f] + r - 1} \tag{15}$$

where r is the index of the customers and $r \in [1, n_{jk}^f]$, $p(T_{jkr}^f = 1)$ denotes the probability of that the r-th customer chooses a new table at which to eat dish k in the restaurant j of the area f. Hence we can sample every T_{jkr} from this Bernoulli distribution on the basis of Eq. (15), then the T_{jk}^f can be computed by $T_{jk}^f = \sum_r T_{jkr}^f$, and $T_{\cdot k}^f = \sum_j T_{jk}^f$.

By following this strategy, the M_k^f can be estimated with:

$$p(M_{kt}^f = 1) = \frac{\gamma^f v_k}{\gamma^f v_k + t - 1} \tag{16}$$

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

Then, we apply a simple multinomial distribution to estimate $T_{fk}^{j\to j+1}$ and $T_{fk}^{0\to j+1}$ within $T_{(j+1)k}^f$. The steps is given below:

$$(T_{fk}^{j\to j+1}, T_{fk}^{0\to j+1}) \sim Multinomial(T_{(j+1)k}^f, [p, 1-p]) \tag{17}$$

$$p = \frac{w_j^f \pi_{jk}^f}{(1 - w_j^f)\beta_k^f + w_j^f \pi_{jk}^f} \tag{18}$$

note that the probability p is affected by the component weighting parameters β_k^f and π_{jk}^f.

Next, we will discuss how to sample the topic of visual words and speech words. According to the theory of HDP-LDA, first, the topic of visual word z_{ji}^f can be drawn from the following steps:

$$\begin{aligned} p(z_{ji}^f = k | x_{ji}^f, \ldots) &\propto p(z_{ji}^f = k | \boldsymbol{\pi}_{\boldsymbol{j}}^{\boldsymbol{f}}) p(x_{ji}^f | z_{ji}^f = k, \ldots) \\ &= \begin{cases} \pi_{jk}^f f_k^{-x_{ji}^f}(x_{ji}^f) & \text{if } k \text{ is used} \\ \pi_{ju}^f f_{k^{new}}^{-x_{ji}^f}(x_{ji}^f) & \text{if } k \text{ is new} \end{cases} \end{aligned} \tag{19}$$

if the topic k has been previously used, the $f_k^{-x_{ji}^f}(x_{ji}^f)$ can be expressed with:

$$\begin{aligned} f_k^{-x_{ji}^f}(x_{ji}^f = v) &= \int f(x_{ji}^f = v | \phi_k) p(\phi_k | \boldsymbol{X}_{\boldsymbol{k}}^{-\boldsymbol{f},\boldsymbol{ji}}, H)\, d\phi_k \\ &= \frac{n_{kv}^{-ji} + \tau}{n_{k\cdot}^{-ji} + V\tau} \end{aligned} \tag{20}$$

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

$$\begin{aligned} f_{k^{new}}^{-x_{ji}^f}(x_{ji}^f = v) &= \int f(x_{ji}^f = v | \phi_k) p(\phi_k | H)\, d\phi_k \\ &= \frac{1}{V} \end{aligned} \tag{21}$$

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

Then, depending on the topic collection of visual words $\boldsymbol{Z}_{\boldsymbol{j}}^{\boldsymbol{f}}$, the topic of speech word $\hat{z}_{j\hat{i}}^f$ can be sampled by:

$$\begin{aligned} p(\hat{z}_{j\hat{i}}^f = k | \hat{x}_{j\hat{i}}^f, \ldots) &\propto p(\hat{z}_{j\hat{i}}^f = k | \boldsymbol{Z}_{\boldsymbol{j}}^{\boldsymbol{f}}) p(\hat{x}_{j\hat{i}}^f | \hat{z}_{j\hat{i}}^f = k, \ldots) \\ &= \frac{n_{jk}^f}{n_{j\cdot}^f} \hat{f}_k^{-\hat{x}_{j\hat{i}}^f}(\hat{x}_{j\hat{i}}^f) \end{aligned} \tag{22}$$

Algorithm 1: Cascaded Gibbs Sampling

Initialization;
for $(f = 1;\ f \le F;\ f{+}{+})$ **do**
  for $(j = J^f;\ j \ge 1;\ j{-}{-})$ **do**
    Sample visual words $\boldsymbol{Z}_j^f$;
    Sample speech words $\hat{\boldsymbol{Z}}_j^f$;
    Sample T_{jk}^f, $T_{fk}^{j-1\to j}$ and $T_{fk}^{0\to j}$;
  end
  Sample M_k^f;
end
Sample concentration parameters ξ, γ^f and α_0^f;
Sample $\boldsymbol{\nu}$;
for $(f = 1;\ f \le F;\ f{+}{+})$ **do**
  Sample $\boldsymbol{\beta}^f$;
  for $(j = 1;\ j \le J^f;\ j{+}{+})$ **do**
    Sample $\boldsymbol{\pi}_j^f$;
  end
end

and the $\hat{f}_k^{-\hat{x}_{j\hat{i}}^f}(\hat{x}_{j\hat{i}}^f)$ can be expressed with:

$$\hat{f}_k^{-\hat{x}_{j\hat{i}}^f}(\hat{x}_{j\hat{i}}^f = \hat{v}) = \int \hat{f}(\hat{x}_{j\hat{i}}^f = \hat{v}|\hat{\phi}_k) p(\hat{\phi}_k|\hat{\boldsymbol{X}}_{\boldsymbol{k}}^{-f,j\hat{i}}, \hat{H})\, d\hat{\phi}_k = \frac{\hat{n}_{k\hat{v}}^{-j\hat{i}} + \hat{\tau}}{\hat{n}_{k\cdot}^{-j\hat{i}} + \hat{V}\hat{\tau}} \tag{23}$$

where $\hat{v}$ denotes the index of speech words in terms of the vocabulary, and $n_{j\cdot}^f$ is the marginal form for the counts of visual words as $n_{j\cdot}^f = \sum_k n_{jk}^f$. Similarly, $\hat{n}_{kv}^{-j\hat{i}}$ denotes the counts of the v-th speech word assigned to topic k except for the current one and $\hat{n}_{k\cdot}^{-j\hat{i}} = \sum_{\hat{v}} \hat{n}_{k\hat{v}}^{-j\hat{i}}$, $\hat{V}$ is the total number of speech words in the vocabulary. In addition, $\hat{\tau}$ is a controlling parameter for speech words and we define $\hat{H} = Dir(\hat{\tau})$.

3.4.3 Gibbs Sampling Implementation

Here, we employ a cascaded Gibbs sampler for implementing the posterior inference of Seq-cHDP. On the basis of the sampling procedures of the variables described in Sect. 3.4.2, we provide the pseudo-code for this cascaded Gibbs sampling task in Algorithm 1.

In this algorithm, we need to pay attention because the number of topics k may increase if a new 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. The updating method is specified in [15]. In fact, we just show only one iteration of the whole Gibbs sampling task. It may operate thousands of iterations to obtain all the variables converged depending on the experimental data.

4. Experiments on Video Data

In this section, we will describe the experimental setup and data features in detail and analyze parameter tuning for the experimental system. Finally, the performances for the proposed and baseline models will be evaluated and compared.

4.1 Experiment Description

The source video documents utilized in this experiment were originally collected from the "blip.tv" video hosting service [17], and processed by *MediaEval-2011 Genre Tagging Task*[†]. In their task, they extracted a series of key frames for each video file, and tagged each key frame with a number of speech transcript words by the automatic speech recognition (ASR) method [17].

Next, we use SIFT descriptor [18] to extract visual words from each key frame [5]. The SIFT descriptor for every 10×10 pixel grid in each key frame is computed conditioning in which the patch size is randomly sampled between scales of 10 to 30 pixels. By using a k-means algorithm, all the obtained SIFT descriptors are clustered into k clusters, which are treated as visual words with a vocabulary size of k.

Table 1 shows quantitative descriptions for experimental video data. Each video file has a genre label attached. We set the vocabulary size of visual words to 1000. For the original speech transcript words, 418 types of standard stop words [19] and the speech words appearing in fewer than five video files are removed, and then the other 6291 types of common words are utilized for the simulation.

We perform genre classification to evaluate the target model. In this process, a nested cross-validation mechanism is employed in our experiment, where the video datasets are evenly divided into five subsets for the use of cross-validation. One subset is treated as the test set while the remaining four subsets are used for four-fold cross-validation. This procedure is totally repeated five times. During the testing phase, the system estimates the genre tag for each video file in the test set based on the training results by employing support vector machines (SVM)[††]. We apply accuracy (micro-F1) and macro-F1 as performance metrics for evaluating our model.

Table 1 Quantitative description for experimental video data

the number of video files	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

[†]http://www.multimediaeval.org/mediaeval2011/

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

4.2 Parameter Tuning

As mentioned in Sect. 3.4, the controlling parameters w_j^f, τ and $\hat{\tau}$ existing in the Seq-cHDP model play a very important role in modeling optimization. Before evaluating the performance of the target model, we need to determine the most optimal values for these controlling parameters through a parameter tuning approach. For simplicity, we consider all the weights w_j^f are equal to the same parameter w. In addition, we also assume $\tau = \hat{\tau} = \tau'$.

The parameter tuning task is conducted by using the validation dataset, which is separated from the training dataset. According to a *nested* cross-validation experiment, in each round of the four-fold cross validation, we can draw one of the four subsets as the validation set, which indicates we can conduct four different experiments by testing four individual validation sets separately. Hence, throughout all the five rounds in this cross-validation experiment, we can collect 5×4 sets of experimental results.

Then, we respectively sweep the parameter w and τ' in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\{0.1, 0.5, 1.0, 1.5, 2.0\}$ in the experimental stage. All the concentration parameters are drawn from a gamma prior $Ga(1.0, 1.0)$. The cascaded Gibbs sampling system will cease the iteration when it reaches convergence. The parameter tuning results are illustrated in Fig. 5. We can see that both accuracy and macro-F1 results are sensitive for varying the parameter w and τ'. The Seq-cHDP model achieves the best performance when we initialize $w = 0.5$ and $\tau' = 1.0$.

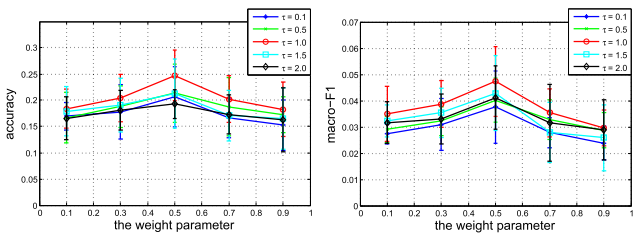


Fig. 5 Result for parameter tuning on Seq-cHDP.

4.3 Evaluation

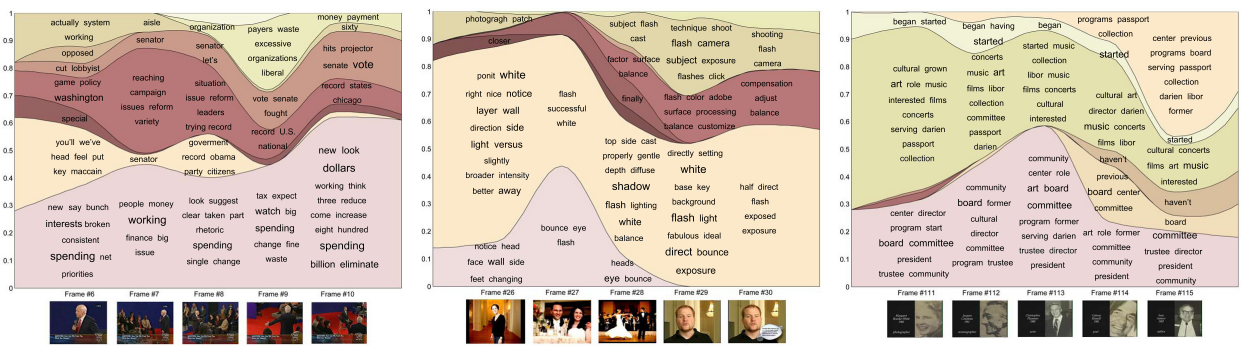
In this section, we present a comprehensive evaluation on Seq-cHDP. For comparison, three types of topic models are taken as baseline models: the first one is a CorrHDP model that is equivalent to the Seq-cHDP model under the condition of $w_j^f = 0$, the second one is a Seq-HDP model that is mentioned in Sec.3.1, and the third one is CorrLDA [4] whose number of topics must be initialized with a fixed positive integer. For parameter settings, we refer to the parameter tuning task given in Sect. 4.2, and set $w = 0.5$ and $\tau' = 1.0$.

4.3.1 Trend Estimation for Latent Topics

Similar to other time dependent topic models [12], [20], one of the advantages of Seq-cHDP is that it can learn and track the trend of the latent topics. However, some of the other baseline topic models fail to learn the varying trend on topic distribution, since they do not have the time-dependency mechanism.

To demonstrate the performance of topic trend estimation, we plot a frame-based area graph of topic distribution on three different examples that occurred in three corresponding video files, shown in Fig. 6. 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 based on 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. Moreover the font size of these key speech words is proportional to their appearance frequencies in each key frame.

According to all these three examples, we see that the topic distribution smoothly varies along with key frame evolving due to the time-dependency mechanism incorporated in Seq-cHDP. Furthermore, the speech words that have close meanings or belong to a similar class are effectively clustered into a specific topic. For example, according to Fig. 6 (a), the real video content is a US presidential can-



(a) Example in video No.106 © Span-TV

(b) Example in video No.109 © Blue Orchid Productions

(c) Example in video No.221 © Darien Library

Fig. 6 Area graph for topic distribution, superimposed with speech words.

Table 2 Performance of genre classification

	Accuracy		Macro-F1	
	Mean	STDEV	Mean	STDEV
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

didate debate on the question "Can we trust you with our money?" In this scenario, we can find that the major topic colored in pink gathers a large number of similar semantic words such as *finance, spending, dollars, tax*, which suggest this topic is probably related to the key word "money". Similarly, for Fig. 6 (b), the video introduced four signature lighting techniques. This explains why the key words *white, flash, shadow, light* often appear in the major topic colored in yellow.

However, the phenomenon of topic fission and fusion for speech words may occur when the topic border of these speech words is not very distinct. For instance in Fig. 6 (c), the speech words *previous, board, center, committee* are separated from the topic colored in pink at Frame #113 and assigned to a new topic colored in orange at Frame #114, since the topic for these speech words is ambiguous.

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

4.3.2 The Performance of Genre Classification

We present the performance of genre classification for Seq-cHDP, CorrHDP, Seq-HDP and CorrLDA in this section. For the CorrHDP model, the controlling parameter τ' is also set to 1. For the Seq-HDP model, the controlling parameter w is also set to 0.5, and the model is respectively evaluated by using single visual data and single speech data. For the CorrLDA model, we respectively conduct three sets of the experiments with different number of topics $K = 10$, $K = 20$ and $K = 40$. As mentioned in Sect. 4.1, all the models are evaluated through a five-fold cross validation four times (with four different systematic random seeds).

The results are shown in Table 2. As we can see, the accuracy and Macro-F1 results of Seq-cHDP achieve 0.2515 and 0.0558 respectively, which outperform those of the other baselines listed 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 video files is imbalanced. In addition, CorrHDP performs at almost the same level as CorrLDA initialized with the best fitting K, hence we infer 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 shows poorer performance than Seq-cHDP under the same condition, which proves the effectiveness of the correspondence method for multimodal data in Seq-cHDP.

5. Conclusions

This paper presents a sequential correspondence hierarchical Dirichlet processes (Seq-cHDP) model to deal with the multimodal data mining issue for video files. Seq-cHDP can be deemed as an extension of HDP that incorporates a time-dependency mechanism and a correspondence method. In experimentation, we evaluated our model by showing the trend estimation for latent topics within a single video file and the performance of genre classification, and finally demonstrated that Seq-cHDP outperforms other baselines in terms of both accuracy and macro-F1.

In future work, we will consider more sophisticated time-dependency mechanism like HMM [10] to enhance the flexibility of our model, and improve our model to achieve higher accuracy. In addition, some supervised topic modeling approaches [21] can also be extended in our model for video classification.

Acknowledgments

This work was supported in part by the Grant-in-Aid for Scientific Research (#15H02703) from JSPS, Japan.

References

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, 2003.
[2] 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, pp.1566–1581, 2006.
[3] Y.W. Teh, "Dirichlet process," in Encyclopedia of Machine Learning, pp.280–287, Springer, 2010.
[4] 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, 2003.
[5] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, pp.524–531, 2005.
[6] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp.1903–1910, 2009.
[7] F. Souvannavong, B. Merialdo, and B. Huet, "Latent semantic analysis for an effective region-based video shot retrieval system," Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, pp.243–250, 2004.
[8] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical Bayesian models," Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pp.1–8, 2007.
[9] T. Hospedales, S. Gong, and T. Xiang, "A Markov clustering topic model for mining behaviour in video," Computer Vision, 2009 IEEE 12th International Conference on, pp.1165–1172, 2009.
[10] D. Kuettel, M.D. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? discovering spatio-temporal dependencies in dynamic scenes," Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp.1951–1958, 2010.
[11] Y. Xie 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] 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, 2010.
[13] T.S. Ferguson, "A Bayesian analysis of some nonparametric problems," The annals of statistics, vol.1, no.2, pp.209–230, 1973.
[14] J. Sethuraman, "A constructive definition of Dirichlet priors," tech. rep., DTIC Document, 1991.
[15] Y.W. Teh and M.I. Jordan, "Hierarchical Bayesian nonparametric models with applications," Bayesian nonparametrics, vol.1, 2010.
[16] I. Porteous, Networks of Mixture Blocks for Non Parametric Bayesian Models with Applications DISSERTATION, Ph.D. thesis, University of California, Irvine, 2010.
[17] 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.
[18] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol.60, no.2, pp.91–110, 2004.
[19] J.P. Callan, W.B. Croft, and S.M. Harding, "The INQUERY retrieval system," Database and Expert Systems Applications, pp.78–83, Springer, 1992.
[20] D.M. Blei and J.D. Lafferty, "Dynamic topic models," Proceedings of the 23rd international conference on Machine learning, pp.113–120, 2006.
[21] J.D. Mcauliffe and D.M. Blei, "Supervised topic models," in Advances in Neural Information Processing Systems, pp.121–128, MIT Press, 2008.

Jianfei Xue is currently pursuing a Ph.D degree at the Graduate School of System Informatics, Kobe University, Japan.

Koji Eguchi is an Associate Professor at the Graduate School of System Informatics, Kobe University, Japan. His research interests include information retrieval, statistical machine learning, and data mining.