IEICE TRANS. INF. & SYST., VOL.E97–D, NO.4 APRIL 2014

PAPER Special Section on Data Engineering and Information Management

Multimedia Topic Models Considering Burstiness of Local Features

Yang XIE[†a)], ***Nonmember and*** **Koji EGUCHI**[†b)], ***Member***

SUMMARY A number of studies have been conducted on topic modeling for various types of data, including text and image data. We focus particularly on the burstiness of the local features in modeling topics within video data in this paper. Burstiness is a phenomenon that is often discussed for text data. The idea is that if a word is used once in a document, it is more likely to be used again within the document. It is also observed in video data; for example, an object or visual word in video data is more likely to appear repeatedly within the same video data. Based on the idea mentioned above, we propose a new topic model, the Correspondence Dirichlet Compound Multinomial LDA (Corr-DCMLDA), which takes into account the burstiness of the local features in video data. The unknown parameters and latent variables in the model are estimated by conducting a collapsed Gibbs sampling and the hyperparameters are estimated by focusing on the fixed-point iterations. We demonstrate through experimentation on the genre classification of social video data that our model works more effectively than several baselines.

key words: *topic models, multimedia, word burstiness, Dirichlet compound multinomials*

1. Introduction

The amount of data worldwide has been explosively increasing because of the widespread use of the Internet and the recent rapid development of social media. These include not only text data but also a lot of images, sounds, and video data. In particular, Internet video took up more than half of all consumer Internet traffic in 2012, and it is forecasted to increase even more in the years to come[†]. However, it is difficult for users to find relevant objects in large-scale video data, and therefore, more sophisticated information access techniques than those currently available are required. Analyzing video via machine learning is one of the emerging research subjects due to the huge cost of manually creating an index of such large-scale video data. Topic modeling approaches such as Latent Dirichlet Allocation (LDA) [1] are particularly promising for tackling the problem mentioned above. Topic models are basically grounded in the idea that each document is generated from a mixture distribution of latent topics, each of which is represented as a multinomial distribution over words. Topic models have already been applied to various data, including text data [1] and image data [2]–[4]. Correspondence LDA (CorrLDA or cLDA) [2] particularly provides a good theory on topic modeling for multimodal data, such as text-annotated images. However, straightforwardly applying CorrLDA to video data is not a good idea since video data usually involves a complex structure, while CorrLDA assumes text-annotated image data, not video data. In this paper, we focus particularly on the burstiness of the local features in modeling the topics within the video data. The burstiness is a phenomenon that is often discussed for text data. Basically, if a word is used once in a document, it is more likely to be used again in the document. This is also observed in video data; for example, an object in video data is more likely to appear in the same video than in different video data. Without taking the burstiness into consideration, the topic models cannot deal with two representations that can be deemed essentially the same but differently appearing in different data, thus disabling the capturing of the diversity in representing the topics. We take into account the burstiness by assuming a different per-topic multinomial distribution over the local features, such as visual words [3], [5], for each video data. Based on the idea mentioned above, we propose a new topic model, Correspondence Dirichlet Compound Multinomial LDA (Corr-DCMLDA), which takes into account the burstiness of the local features in video data. We evaluated our model through experimentation on genre classification, and demonstrate that our model works more effectively than several baselines.

2. Related Work

Some researchers have explored topic models for image data [2]–[4]. CorrLDA [2] particularly provides a good theory for modeling the dependencies between an image and the text features. Whereas CorrLDA is based on the premise that the target is a collection of text-annotated image data, our motivation is to model multimodal video documents consisting of a sequence of key frame images with speech transcripts. Topic models have also been applied to video data [6]–[10]. For instance, Souvannavong et al. [6] extended Probabilistic Latent Semantic Analysis (PLSA) for the task of object retrieval and scene classification using video data. Wanke et al. [9] addressed semantics-preserving video compression using PLSA, and achieved about 2 : 1 compression ratio, compared to other dimension reduction techniques, while maintaining the prediction capability.

Manuscript received July 8, 2013.
Manuscript revised October 29, 2013.

[†]The authors are with the Graduate School of System Informatics, Kobe University, Kobe-shi, 657–8501 Japan.
a) E-mail: xieyang@cs25.scitec.kobe-u.ac.jp
b) E-mail: eguchi@port.kobe-u.ac.jp

DOI: 10.1587/transinf.E97.D.714

[†]Cisco Visual Networking Index: http://www.cisco.com/en/US/netsol/ns827/networking_solutions_solution_category.html

These studies explored a topic modeling that can capture multimodal information; however, they did not consider the burstiness of visual or text features.

On the other hand, we explore whether and to what extent the burstiness of visual and text features have an impact on topic modeling for video data. For this purpose, we model the burstiness of visual and text features in topic modeling for video data. DCMLDA [11], [12] provides appropriate means to take into account the word burstiness in topic modeling. However, DCMLDA is based on the premise that the target is a collection of unimodal (text-only) data. We will describe in more detail the CorrLDA and DCMLDA within the context of modeling multimodal video documents in Sects. 2.1 and 2.2, respectively.

2.1 CorrLDA

CorrLDA [2] is a topic model that was proposed for text-annotated image data to simultaneously model visual features and text words. In this modeling, it first generates topics for the visual features in an annotated image. Then, only the topics associated with the visual features in the image are used to generate text words. In the original CorrLDA, each visual feature is assumed to be generated from a multivariate Gaussian distribution conditioned on a latent topic. The multivariate Gaussian distribution can be replaced by a multinomial distribution with a Dirichlet prior, when we represent the visual features as visual words, such as those used in [3], [5]. We use this replacement in this paper. When we apply CorrLDA to video documents, we need to assume that the entire video collection consists of a set of key frames, disregarding the video documents unit. Figure 2 shows a graphical model representation of CorrLDA, where F, K, N_f, and N'_f respectively indicate the number of key frames, number of topics, and numbers of visual words and

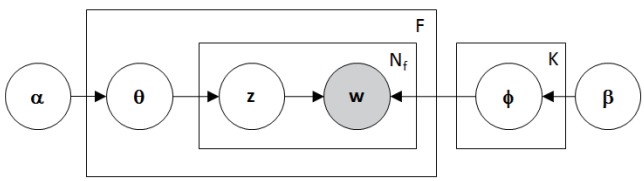


Fig. 1 Graphical model of LDA.

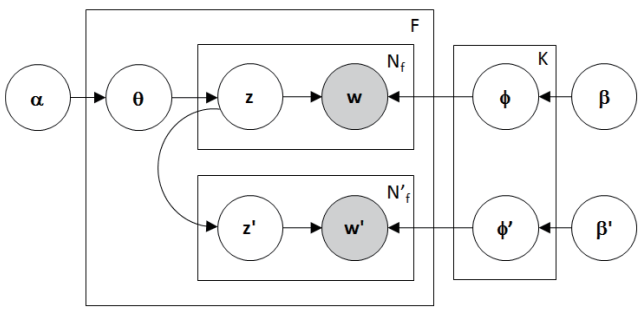


Fig. 2 Graphical model of CorrLDA.

speech transcript words that are associated with key frame f. For reference, Fig. 1 shows a graphical model representation of LDA. CorrLDA's generative process is shown as below, where the prime mark indicates the variables corresponding to speech transcript words:

1. For all key frames f, sample $\boldsymbol{\theta}_f \sim Dirichlet(\boldsymbol{\alpha})$.
2. For all topics k, sample $\boldsymbol{\phi}_k \sim Dirichlet(\boldsymbol{\beta})$ and $\boldsymbol{\phi}'_k \sim Dirichlet(\boldsymbol{\beta}')$.
3. For each of the N_f visual words w_{fi} in key frame f:
   a. Sample a topic $z_{fi} \sim Multinomial(\boldsymbol{\theta}_f)$.
   b. Sample a visual word $w_{fi} \sim Multinomial(\boldsymbol{\phi}_{z_{fi}})$.
4. For each of the N'_f speech transcript words w'_{fi} in key frame f:
   a. Sample a topic $z'_{fi} \sim Uniform(z_{f1}, \cdots, z_{fN_f})$.
   b. Sample a speech transcript word $w'_{fi} \sim Multinomial(\boldsymbol{\phi}'_{z'_{fi}})$.

We estimate the latent variables and unknown parameters of CorrLDA by conducting a collapsed Gibbs sampling. The full conditional probability of generating topic k for visual word n (or speech transcript word n') in key frame f is given by:

$$
\begin{aligned}
&p(z_{f,n} = k | Z_{\neg(f,n)}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&\quad \propto \frac{c(f,k)_{\neg(f,n)} + \alpha_k}{\sum_k c(f,k) + \sum_k \alpha_k} \cdot \frac{c(k,w)_{\neg(f,n)} + \beta_{k,w}}{\sum_w c(k,w)_{\neg(f,n)} + \sum_w \beta_{k,w}} \\
&p(z'_{f,n'} = k | Z'_{\neg(f,n')}, W', \boldsymbol{\beta}') \\
&\quad \propto \frac{c(f,k)}{\sum_k c(f,k)} \cdot \frac{c(k,w')_{\neg(f,n')} + \beta'_{k,w'}}{\sum_{w'} c(k,w')_{\neg(f,n')} + \sum_{w'} \beta'_{k,w'}}.
\end{aligned} \tag{1}
$$

For the first equation above, $W = \{\mathbf{w}_f\}_{f\in\{1,\ldots,F\}}$ and $\mathbf{w}_f = \{\mathbf{w}_{f,n}\}_{n\in\{1,\ldots,N_f\}}$, where $\mathbf{w}_{f,n}$ represents a random variable of visual word n in key frame f. $Z = \{\mathbf{z}_f\}_{f\in\{1,\ldots,F\}}$ and $\mathbf{z}_f = \{\mathbf{z}_{f,n}\}_{n\in\{1,\ldots,N_f\}}$, where $\mathbf{z}_{f,n}$ represents a random variable of a topic assigned to word $\mathbf{w}_{f,n}$. $c(f,k)$ and $c(k,w)$ indicate that the count of topic k is assigned to key frame f and the count of topic k is assigned to visual word n, respectively. The subscript $\neg(f,n)$ indicates the removal of the topic that was previously assigned to word n in key frame f. As for the hyperparameters, $\boldsymbol{\alpha} = \{\alpha_k\}_{k\in\{1,\ldots,K\}}$ and $\boldsymbol{\beta} = \{\beta_w\}_{w\in\{1,\ldots,V\}}$. The second equation in Eq. (1) corresponds to the full conditional probability of generating topic k for speech transcript word n' in key frame f. Here, some notations are specified using the prime mark to distinguish those for speech transcript words. Note that the first term of the right-hand side of the second equation indicates the relative frequency of the topics that are assigned to visual words.

Moreover, we can estimate asymmetric Dirichlet hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by using fixed-point iterations [13] according to:

$$
\begin{aligned}
\alpha_k^{new} &= \alpha_k \cdot \frac{\sum_f \Psi(c(f,k) + \alpha_k) - \Psi(\alpha_k)}{\sum_f \Psi(\sum_k c(f,k) + \sum_k \alpha_k) - \Psi(\sum_k \alpha_k)} \\
\beta_w^{new} &= \beta_w \cdot \frac{\sum_k \Psi(c(k,w) + \beta_w) - \Psi(\beta_w)}{\sum_k \Psi(\sum_w c(k,w) + \sum_w \beta_w) - \Psi(\sum_w \alpha_w)},
\end{aligned}
$$

where $\Psi(\cdot)$ represents a digamma function. Using asymmetric hyperparameter $\boldsymbol{\beta}$ sometimes decreased the model performance in experiments using text data [14]. Then, symmetric Dirichlet hyperparameter $\boldsymbol{\beta} = (\beta, \cdots, \beta)$ with a common β can be used instead. Here after in this paper, CorrLDA with asymmetric hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is referred to as 'CorrLDA-asym', and CorrLDA with symmetric hyperparameter $\boldsymbol{\beta}$ and asymmetric hyperparameter $\boldsymbol{\alpha}$ as 'CorrLDA-sym'.

2.2 DCMLDA

LDA and most of its variants alone cannot capture the word burstiness, a phenomenon that if a word is used once in a document, it is more likely to be used again in the document. Some researchers have incorporated Dirichlet compound multinomials into LDA to capture the word burstiness when modeling digitized books [11] and in modeling documents [12]. Those models are both called DCMLDA (Dirichlet Compound Multinomial LDA). We will briefly describe DCMLDA according to the latter, in the context of modeling video data below.

In modeling video data, DCMLDA can take into consideration that each video document consists of a set of (independent) key frames. We made an assumption that each key frame is represented by a set of mixed discrete features of visual words and speech transcript words, since DCMLDA cannot directly handle such multimodal data. Figure 3 shows a graphical model representation of DCMLDA, where D, K, F_d, and N_{df} respectively represent the number of video documents, number of topics, number of key frames in video document d, and number of discrete features that are associated with key frame f of video document d. The generative process of DCMLDA can be shown as follows.

1. For all key frames f in each of all video documents d, sample $\boldsymbol{\theta}_{df} \sim Dirichlet(\boldsymbol{\alpha}_d)$.
2. For all topics k and for all video documents d, sample $\boldsymbol{\phi}_{dk} \sim Dirichlet(\boldsymbol{\beta}_k)$.
3. For each of the N_{df} discrete features w_{dfi} in key frame f of video document d:
   a. Sample topic $z_{dfi} \sim Multinomial(\boldsymbol{\theta}_{df})$.
   b. Sample discrete feature $w_{dfi} \sim Multinomial(\boldsymbol{\phi}_{dz_{dfi}})$.

Note that DCMLDA disregards the time dependency of the key frames within each video document.

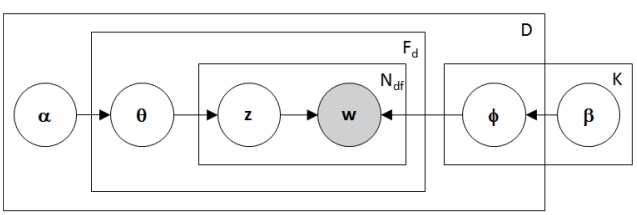


Fig. 3 Graphical model of DCMLDA.

3. Corr-DCMLDA

3.1 Formalization

We focus on the burstiness of the local features in video documents, where a visual word or speech transcript word is more likely to appear repeatedly within the same video document. CorrLDA can capture the dependencies between visual and text words in text-annotated image collections; however, when we apply CorrLDA to video documents, it deems the video documents to be a set of key frames, disregarding the video documents unit. Moreover, CorrLDA is unable to do anything about the burstiness of these features. DCMLDA can take into consideration that each video document consists of a set of (independent) key frames, and furthermore, it can take into account the burstiness. However, DCMLDA cannot appropriately capture the dependencies between the visual words and speech transcript words. We propose Corr-DCMLDA, whose graphical model is given in Fig. 4 and whose generative process is given as follows to address the problems.

1. For all key frames f in each of all video documents d, sample $\boldsymbol{\theta}_{df} \sim Dirichlet(\boldsymbol{\alpha}_d)$.
2. For all topics k and for all video documents d, sample $\boldsymbol{\phi}_{dk} \sim Dirichlet(\boldsymbol{\beta}_k)$ and $\boldsymbol{\phi}'_{dk} \sim Dirichlet(\boldsymbol{\beta}'_k)$.
3. For each of the N_{df} visual words w_{dfi} in key frame f of video document d:
   a. Sample topic $z_{dfi} \sim Multinomial(\boldsymbol{\theta}_{df})$.
   b. Sample visual word $w_{dfi} \sim Multinomial(\boldsymbol{\phi}_{dz_{dfi}})$.
4. For each of the N'_{df} speech transcript words w'_{dfi} in key frame f of video document d:
   a. Sample topic $z'_{dfi} \sim Uniform\left(z_{df1}, \cdots, z_{dfN_{df}}\right)$.
   b. Sample speech transcript word $w'_{dfi} \sim Multinomial(\boldsymbol{\phi}'_{dz'_{dfi}})$.

We present the flow for the extraction of features and the estimation of unknown parameters when using Corr-DCMLDA in Fig. 5. We will describe the details in the rest of this section.

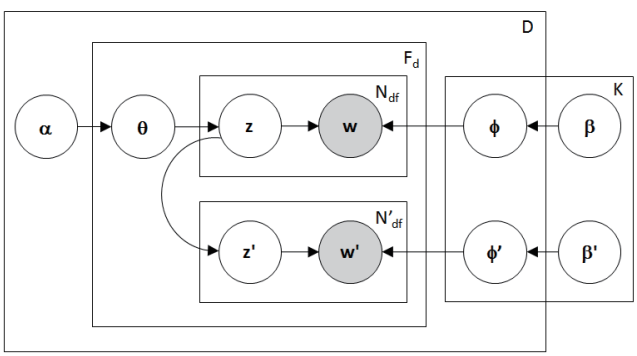


Fig. 4 Graphical model of Corr-DCMLDA.

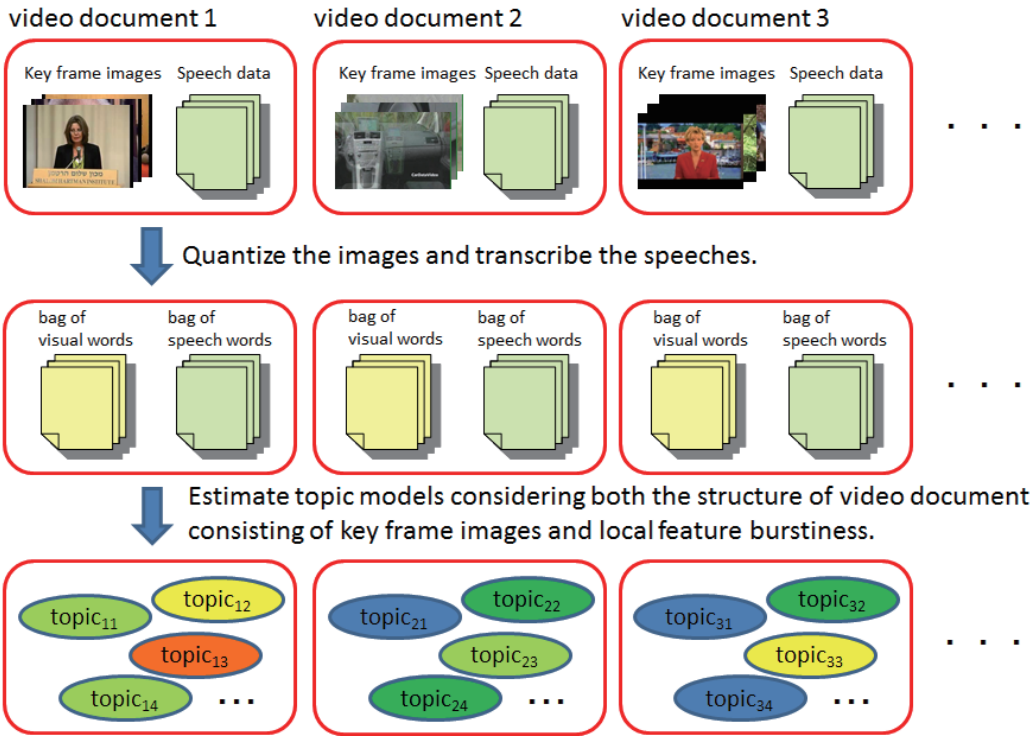


Fig. 5 Flow for estimating Corr-DCMLDA.

3.2 Features

We use visual words and speech transcript words as the features of the key frames of each video document. We first compute the SIFT descriptor [15] for every 10×10-pixel grid of each key frame, assuming that the size of the patch is randomly sampled between scale 10 to 30 pixels [3] to extract visual words. The resulting SIFT descriptors are then clustered using a k-means algorithm, and the resulting k clusters are used as visual words with a vocabulary size of k.

3.3 Estimation

We estimate the latent variables and unknown parameters of Corr-DCMLDA by conducting a collapsed Gibbs sampling. The full conditional probability of generating a topic k for visual word n (and speech transcript word n') in key frame f of video document d is given by:

$$
\begin{aligned}
&p(z_{d,f,n} = k | Z_{\neg(d,f,n)}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&\quad \propto \frac{c(d,f,k)_{\neg(d,f,n)} + \alpha_{d,k}}{\sum_k c(d,f,k) + \sum_k \alpha_{d,k}} \cdot \frac{c(d,k,w)_{\neg(d,f,n)} + \beta_{k,w}}{\sum_w c(d,k,w)_{\neg(d,f,n)} + \sum_w \beta_{k,w}} \\
&p(z'_{d,f,n'} = k | Z'_{\neg(d,f,n')}, W', \boldsymbol{\beta}') \\
&\quad \propto \frac{c(d,f,k)}{\sum_k c(d,f,k)} \cdot \frac{c(d,k,w')_{\neg(d,f,n')} + \beta'_{k,w'}}{\sum_{w'} c(d,k,w')_{\neg(d,f,n')} + \sum_{w'} \beta'_{k,w'}}.
\end{aligned} \tag{2}
$$

For the first equation above, $W = \{\mathbf{w}_{d,f}\}_{d\in\{1,\ldots,D\},f\in\{1,\ldots,F\}}$ and $\mathbf{w}_{d,f} = \{\mathbf{w}_{d,f,n}\}_{n\in\{1,\ldots,N_{d,f}\}}$, where $\mathbf{w}_{d,f,n}$ represents a random variable of visual word n in key frame f of video document d. $Z = \{\mathbf{z}_{d,f}\}_{d\in\{1,\ldots,D\},f\in\{1,\ldots,F\}}$ and $\mathbf{z}_{d,f} = \{\mathbf{z}_{d,f,n}\}_{n\in\{1,\ldots,N_{d,f}\}}$, where $\mathbf{z}_{d,f,n}$ represents a random variable of a topic assigned to word $\mathbf{w}_{d,f,n}$. $c(d,f,k)$ and $c(d,k,w)$ represent the count of topic k that is assigned to key frame f in video document d and the count of topic k that is assigned to visual word n in video document d, respectively. The subscript $\neg(d,f,n)$ represents the removal of the topic that was previously assigned to word n in key frame f of video document d. As for the hyperparameters, $\boldsymbol{\alpha} = \{\alpha_{d,k}\}_{d\in\{1,\ldots,D\},k\in\{1,\ldots,K\}}$ and $\boldsymbol{\beta} = \{\beta_{k,w}\}_{d\in\{1,\ldots,D\},w\in\{1,\ldots,V\}}$. The second equation in Eq. (2) corresponds to the full conditional probability of generating topic k for speech transcript word n' in key frame f of video document d. Here, some notations are specified by the prime mark to distinguish those for speech transcript words. Note that the first term of the right-hand side in the second equation represents the relative frequency of the topics that are assigned to visual words, as in CorrLDA [2].

Moreover, we can estimate hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by using fixed-point iterations [13] according to:

$$
\begin{aligned}
\alpha^{new}_{d,k} &= \alpha_{d,k} \cdot \frac{\sum_f \Psi(c(d,f,k) + \alpha_{d,k}) - \Psi(\alpha_{d,k})}{\sum_f \Psi(\sum_k c(d,f,k) + \sum_k \alpha_{d,k}) - \Psi(\sum_k \alpha_{d,k})} \\
\beta^{new}_{k,w} &= \beta_{k,w} \cdot \frac{\sum_d \Psi(c(d,k,w) + \beta_{k,w}) - \Psi(\beta_{k,w})}{\sum_d \Psi(\sum_w c(d,k,w) + \sum_w \beta_{k,w}) - \Psi(\sum_w \alpha_{k,w})},
\end{aligned}
$$

where $\Psi(\cdot)$ represents a digamma function.

Table 1 Summary of dataset used.

No. of video documents	132
No. of key frames	4596
Vocabulary size of visual words	1000
Vocabulary size of speech transcript words	6291
No. of genre classes	13

Table 2 List of genres used in dataset.

genre code	genre name	no. of video documents
1002	business	7
1003	citizen_journalism	11
1004	comedy	6
1008	educational	19
1011	health	7
1012	literature	6
1013	movies_and_television	6
1015	personal_or_auto-biographical	5
1017	religion	16
1019	sports	7
1020	technology	27
1022	the_mainstream_media	5
1024	videoblogging	10

4. Experiments

4.1 Data

We used a video document collection that was developed in the MediaEval-2011 Tagging Task[†] and originally collected from the "blip.tv" video hosting service [16]. A summary of the dataset used is listed in Table 1. Each video document in this dataset is associated with a genre label, as shown in Table 2, and therefore, we evaluate the models in the genre classification task. Note that we removed, from the original dataset, 'politics' and 'default_category' that make the dataset too imbalanced and some other genres that are associated with less than five video documents. As the result, the number of topics is 13 while that was 26 in the original dataset.

Each video document consists of the key frames associated with speech transcripts. Each key frame is extracted from the middle of the sequence of the frames for each shot, which was automatically segmented by [17]. We extracted visual words from each key frame image in the manner described in Sect. 3.2. We set the number of visual words to 1000, according to our preliminary experiments. We removed 418 types of standard stop words [18] from the speech transcript words. We also removed the speech words that occurred in less than five video documents.

4.2 Evaluation

We evaluate the models in the genre classification task. We randomly split the dataset into five subsets, and then we retain one single subset as a test set. Using the remaining four subsets, we determine two free parameters (the number of topics and regularization parameter that we will mention later in this section) for each model via 4-fold cross-validation ('cross-validation stage'). We finally evaluate each model using the test set ('test stage').

For Corr-DCMLDA, we estimate the latent variables and unknown parameters using the collapsed Gibbs sampling and estimate asymmetric Dirichlet hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via a fixed-point iteration, as mentioned in Sect. 3.3. For held-out data, we estimate latent variables of topic assignments Z and Z' using $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ that were estimated previously. We use CorrLDA as a baseline model assuming that the entire video collection consists of a set of key frames and by disregarding the unit of video documents. We estimate the latent variables and unknown parameters in CorrLDA using a collapsed Gibbs sampling and estimate asymmetric Dirichlet hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via a fixed-point iteration ('CorrLDA-asym'), as mentioned in Sect. 2.1. As an alternative baseline, we set symmetric hyperparameter $\boldsymbol{\beta}$ at 0.1 and estimate asymmetric hyperparameter $\boldsymbol{\alpha}$ ('CorrLDA-sym'). For held-out data, we estimate Z and Z' using $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$ that were estimated previously, in the case of both baselines.

We then learn logistic regression model [19] as a classifier using $\bar{\mathbf{z}}_{d,f} = \mathbb{E}_{n\in\{1,\ldots,N_{d,f}\}}[\mathbf{z}_{d,f,n}]$ and $\bar{\mathbf{z}}'_{d,f} = \mathbb{E}_{n'\in\{1,\ldots,N'_{d,f}\}}[\mathbf{z}'_{d,f,n'}]$ as explanatory variables[††]. If we formally describe it, the probability of key frame f in video document d falling into genre y can be given by:

$$p(y|\bar{\mathbf{z}}_{d,f}, \bar{\mathbf{z}}'_{d,f}, \{\boldsymbol{\eta}^{(c)}\}_{c\in\mathcal{Y}}) = \frac{\exp\left(\boldsymbol{\eta}^{(y)\mathrm{T}}(\bar{\mathbf{z}}_{d,f}, \bar{\mathbf{z}}'_{d,f})\right)}{\sum_{c\in\mathcal{Y}}\exp\left(\boldsymbol{\eta}^{(c)\mathrm{T}}(\bar{\mathbf{z}}_{d,f}, \bar{\mathbf{z}}'_{d,f})\right)},$$

where $\mathcal{Y}$ represents a set of genres. $(\bar{\mathbf{z}}_{d,f}, \bar{\mathbf{z}}'_{d,f})$ represents the concatenation of the two K-dimensional vectors. $\boldsymbol{\eta}^{(y)}$ indicates a $2K$-dimensional weight vector corresponding to genre y. To avoid overfitting, we employ L_2 regularization with parameter C, which is determined by cross-validation. The weight vector is then learned by using the L-BFGS Quasi-Newton algorithm [21]. As we previously mentioned, each video document is associated with a genre label, as listed in Table 2. Therefore, we assume that all the key frames are associated with the genre label given with each video document, in the process of classifier learning. For class prediction, we take a genre label of video document d as $\mathbf{argmax}_{y\in\mathcal{Y}}\mathbb{E}_{f\in\{1,\ldots,N_d\}}[p(y|\bar{\mathbf{z}}_{d,f}, \bar{\mathbf{z}}'_{d,f}, \{\boldsymbol{\eta}^{(c)}\}_{c\in\mathcal{Y}})]$.

In Fig. 6, we present the accuracy of Corr-DCMLDA, CorrLDA-asym, and CorrLDA-sym, varying the number of topics (as $K = \{5, 10, 20, 40, 80\}$) and the regularization parameter (as $C = \{0.01, 0.1, 1, 10, 100\}$), where the resulting accuracy was averaged over the four splits in the process of 4-fold cross-validation. Table 3 indicates the optimal parameters determined by this process. Table 4 shows the test results in terms of accuracy and macro-F1. As you can see in Fig. 6 and Table 3, the accuracy is sensitive to both the number of topics K and the regularization parameter C, and

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

[††]We used "classias" [20] for logistic regression.

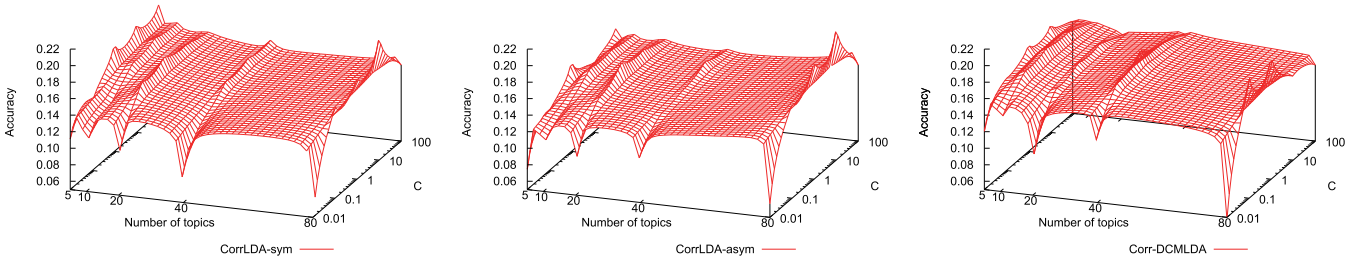


Fig. 6 Accuracy in genre classification at cross-validation stage.

Table 3 Optimal parameters of the number of topics K and the regularization parameter C.

	K	C
CorrLDA-sym	10	10
CorrLDA-asym	80	10
Corr-DCMLDA	40	1

Table 4 Accuracy and macro-F1 in genre classification at test stage.

	Accuracy	Macro-F1
CorrLDA-sym	0.1482	0.0228
CorrLDA-asym	0.1852	0.0256
Corr-DCMLDA	0.2813	0.0547
	(+0.8984)	(+1.3999)
	[+0.5188]	[+1.1332]

(·) and [·] indicate the rate of improvement achieved by Corr-DCMLDA over CorrLDA-sym and CorrLDA-asym, respectively.

moreover, the optimal K and C are different for each model. The test results in Table 4 show that Corr-DCMLDA outperformed CorrLDA-sym and CorrLDA-asym in terms of both accuracy and macro-F1 in the task of genre classification, and therefore, this supports the notion that expressiveness of Corr-DCMLDA is more powerful than that of CorrLDA. From the comparison between CorrLDA-asym and CorrLDA-sym, we found that using asymmetric Dirichlet hyperparameter β does not damage the model performance in the experiments, which is different from that reported in the experiments using text data [14]. This result is consistent with our intuition that the estimation of β is more successful when using CorrLDA for image data associated with text data than LDA for text data, probably because the size of the visual words vocabulary is smaller than that of the text words and the topic assignments of visual words are more dominant than those of text words, when applying CorrLDA to such data.

5. Conclusions

We focused on the burstiness of the local features in the modeling topics within video data, and proposed Corr-DCMLDA for this purpose, which is an extension of CorrLDA that provides a good theory on topic modeling for multimodal data. We evaluated Corr-DCMLDA through experimentation on the task of genre classification of video documents that consist of visual words and speech transcript words, and demonstrated that Corr-DCMLDA works more effectively than CorrLDA in terms of both accuracy and macro-F1.

A more detailed evaluation is left for our further work. Moreover, incorporating the ideas of Symmetric correspondence topic modeling [22] is promising. It can capture the bidirectional dependency between multiple modes: image features and speech transcript words, while CorrLDA only captures the unidirectional dependency such as from the image features to speech transcript words. Another direction for our future work is modeling video documents associated with social tags and/or social networks in a more sophisticated way. We are planning to incorporate the ideas of supervised topic modeling [23]–[25] into our multimedia topic models.

Acknowledgments

We thank Kiyoshi Nishihara for helping with data handling. This work was supported in part by the Grant-in-Aid for Scientific Research (#23300039) from JSPS, Japan, and in part by Hoso Bunka Foundation.

References

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.
[2] 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, Toronto, Canada, 2003.
[3] F.F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp.524–531, Washington, DC, USA, 2005.
[4] C. Wang, D.M. Blei, and F.F. Li, "Simultaneous image classification and annotation," Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp.1903–1910, Miami, Florida, USA, 2009.
[5] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," ECCV International Workshop on Statistical Learning in Computer Vision, pp.1–22, Prague, Czech Republic, 2004.
[6] 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, New York, NY, USA, 2004.
[7] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception

by hierarchical bayesian models," Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07), pp.1–8, Minneapolis, Minnesota, USA, 2007.

[8] T. Hospedales, S. Gong, and T. Xiang, "A Markov clustering topic model for mining behaviour in video," Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV 2009), pp.1165–1172, Kyoto, Japan, 2009.

[9] J. Wanke, A. Ulges, C.H. Lampert, and T.M. Breuel, "Topic models for semantics-preserving video compression," Proceedings of the 11th International Conference on Multimedia Information Retrieval (MIR'10), pp.275–284, Philadelphia, Pennsylvania, USA, 2010.

[10] P. Das, C. Xu, R.F. Doell, and J.J. Corso, "A thousand frames in just a fewwords: Lingual description of videos through latent topics and sparse object stitching," Proceedings of the 2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2013), pp.2339–2346, Portland, Oregon, USA, 2013.

[11] D. Mimno and A. McCallum, "Organizing the oca: Learning faceted subjects from a library of digital books," Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp.376–385, Vancouver, BC, Canada, 2007.

[12] G. Doyle and C. Elkan, "Accounting for burstiness in topic models," Proceedings of the 26th Annual International Conference on Machine Learning, pp.281–288, Montreal, Quebec, Canada, 2009.

[13] T. Minka, "Estimating a Dirichlet distribution," tech. rep., 2000.

[14] H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," Proceedings of the 26th International Conference on Machine Learning, pp.1105–1112, Montreal, Canada, 2009.

[15] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol.60, no.2, pp.91–110, 2004.

[16] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedl, R. Ordelman, and G.J.F. Jones, "Automatic tagging and geotagging in video collections and communities," Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pp.51:1–51:8, Pisa, Italy, 2011.

[17] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequences," Proceedings of the 10th Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'09), pp.25–28, London, UK, 2009.

[18] J.P. Callan, W.B. Croft, and S.M. Harding, "The INQUERY retrieval system," Proceedings of the 3rd International Conference on Database and Expert Systems Applications, pp.78–83, Valencia, Spain, 1992.

[19] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[20] N. Okazaki, "Classias: A collection of machine-learning algorithms for classification," http://www.chokkan.org/software/classias/, 2009.

[21] J. Nocedal and S.J. Wright, Numerical Optimization, Springer, 2006.

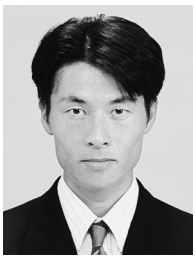
[22] K. Fukumasu, K. Eguchi, and E.P. Xing, "Symmetric correspondence topic models for multilingual text analysis," Advances in Neural Information Processing Systems, pp.1295–1303, 2012.

[23] D. Blei and J.D. McAuliffe, "Supervised topic models," Advances in Neural Information Processing Systems, pp.121–128, 2007.

[24] J. Zhu, A. Ahmed, and E.P. Xing, "MedLDA: Maximum margin supervised topic models for regression and classification," Proceedings of the 26th Annual International Conference on Machine Learning, pp.1257–1264, Montreal, Quebec, Canada, 2009.

[25] J. Zhu, N. Chen, H. Perkins, and B. Zhang, "Gibbs max-margin topic models with fast sampling algorithms," Proceedings of the 30th Annual International Conference on Machine Learning, pp.124–132, Atlanta, Georgia, USA, 2013.

Yang Xie is currently pursuing a Master's 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.