A BAYESIAN NONPARAMETRIC MULTIMODAL DATA MODELING FRAMEWORK FOR VIDEO EMOTION RECOGNITION*

Jianfei Xue, Zhaojie Luo†, Koji Eguchi, Tetsuya Takiguchi, and Tsukasa Omoto‡

Graduate School of System Informatics, Kobe University, Kobe, Hyogo, Japan
xjfxjf008@gmail.com, luozhaojie@me.cs.scitec.kobe-u.ac.jp, eguchi@port.kobe-u.ac.jp,
takigu@kobe-u.ac.jp, omoto@cs25.scitec.kobe-u.ac.jp

ABSTRACT

Video emotion recognition as an emerging research field has been attracting more and more focus in recent years. However, such work is quite challenging, since human emotions are hard to differentiate precisely due to its complexity and diversity, moreover, the expressions of sentiment in a content-rich video are sparse. Previous studies presented a number of approaches to try to learn human emotions on video level by exploiting various video features. However, most of works just used simple low-level video features such as hand-crafted image features, and they also did not consider the further latent connections among different multimodal data within a video. To tackle these problems, we develop a novel Bayesian nonparametric multimodal data modeling framework to learn the emotions from video, where the adopted image data are deep features extracted from key frames of video via convolutional neural networks (CNNs), and the adopted audio data are Mel-frequency cepstral coefficient (MFCC) features. In this framework, we then use a symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) model to mine their latent emotional events (topics) between image features and audio features. Finally, the effectiveness of our framework is demonstrated via comprehensive experimentations.

Index Terms— convolutional neural networks, Bayesian nonparametric methods, emotion recognition

1. INTRODUCTION

Nowadays, people are not merely content with some studies about simple pattern recognition for video. In other words, they are becoming more interested in video deep understanding through some cutting-edge techniques in the field of computer vision and machine learning. For instance, emotion recognition for videos is one of the popular research topics currently. This technique can help us understand the emotion of people shown in a video clip by using visual information and audio information. It is promising to be applied in video recommendation services as auxiliary means, which can effectively find users' interests and recommend the corresponding videos to them based on obtained video emotion.

According to the theories about close interaction between cognitive processes and emotional appraisals [1], human emotions are complex and diverse. For learning emotions from videos, it is a more challenging work, since the expressions of sentiment are sparsely distributed in a video, furthermore, the multimodal data (such as image data and audio data extracted from video) processing and modeling for video emotion recognition are tricky as well. Previous studies provided a number of proposals for emotion recognition in videos. Kang [2] proposed a method for detecting affective events through hidden Markov models (HMM), where simple low-level video features including color, motion and shot cut rate are extracted and utilized for mapping to high-level emotional events via an empirical study. Xu et al. [3], Teixeira et al. [4] and Jiang et al. [5] focused on emotion classification for videos by jointly using both visual features and audio features, however, they just used simple multimodal data fusion methods without considering the further latent connections over all the multimodal data types, moreover, most of adopted visual features are also low-level features. More recently, along with the rapid development of convolutional neural networks (CNNs) [6, 7, 8], people attempt to further improve the performance of emotion recognition via CNNs. Chen et al. [9] and You et al. [10] exploited corresponding CNN models to analyze emotions on image level, the results demonstrated such deep features outperform hand-crafted low-level features and features from SentiBank. Then, Xu et al. [11] first proposed an video emotion recognition framework based on deep features extracted from CNNs, their study showed a comprehensive discussion on the evaluations of emotion recognition among different CNNs and also the features from different layers of each CNN. However, auxiliary images are necessary for deep feature transfer encoding, and the multimodal data fusion strategy in their work also looks simple.

To tackle the problems mentioned above, we present a novel Bayesian nonparametric multimodal data modeling

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

†Jianfei Xue and Zhaojie Luo are considered as equal first authors.

‡Tsukasa Omoto is now with DWANGO Co., Ltd.

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

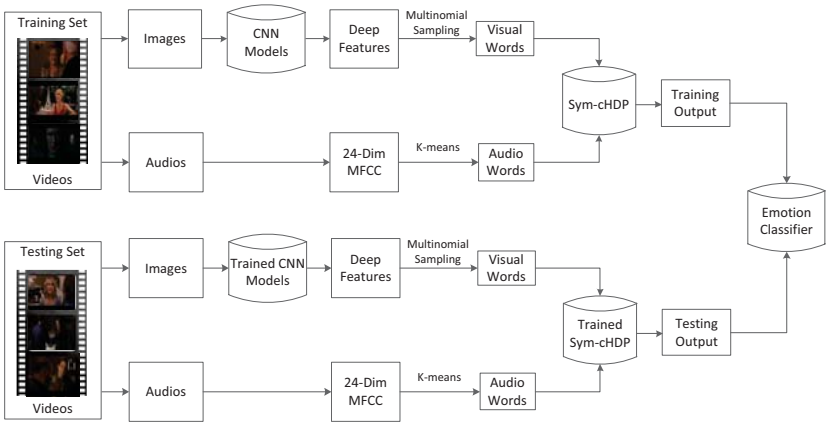

Fig. 1. A flowchart of our proposed method for video emotion classification.

framework for video emotion recognition in this paper. According to the framework, CNN-based deep features and MFCC features drawn from videos are deemed as image features and audio features. Then, we introduce a symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) model [12] to further learn the latent emotional events (topics) from image features and audio features. Based on learning results, a supervised classifier will be utilized to predict an emotional label for each video. Comparative experiments are conducted to evaluate the effectiveness of our method from different aspects.

2. FRAMEWORK

As we mentioned in Sec. 1, human emotion recognition for videos is a quite challenging work due to its complexity, diversity and sparsity. To improve the emotion learning performance, we adopt a Bayesian nonparametric multimodal data modeling method to further understand the latent sentimental information from extracted video deep features.

Firstly, we give an overview of our framework, which is depicted in Fig. 1. From this figure, we can find that there are two core models embedded in this framework: one is a CNN model for deeply learning the images drawn from videos, the other is a Bayesian nonparametric topic model (*i.e.*, Sym-cHDP) for modeling image features and audio features via mining their latent emotional events (topics).

According to the flowchart of our method, image data and audio data are firstly extracted from each video. For images, we then input them to a CNN model. Generally, CNN architecture involving a sequence of deep learning layers is directly utilized for image classification [6, 7, 8]. However, recent study [11] shows that the features drawn from deeper layers in CNN express more in-depth information, such as sentimental information. In our work, we therefore draw the corresponding image deep features via CNN for emotion recognition. The details about CNN models and deep features adopted in our work will be specified in Sec. 3.1. Meanwhile, for audio data, we uniformly sample a 24-dimensional MFCC descriptor 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.

Then, we use a symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) model [12] to handle the multimodal data modeling issue in our framework. Sym-cHDP as an extension of HDP [13] is also a Bayesian nonparametric topic model. Original HDP is well-known for text mining, whose theory assumes that each text document is represented as a mixture of latent topics, and each latent topic is represented as a word distribution. HDP models multiple text documents as multiple infinite Dirichlet processes connected by sharing the same mixture of components (topics). In the case of video emotion analysis, we are inspired that the structure of emotion expression in video multimodal data has a similar form. We respectively use visual words and audio words to represent CNN-based deep features and MFCC features. For each video, the video emotion can be expressed with a collection of latent emotional events occurring in the video, while each emotional event reflects in corresponding visual words and audio words. For instance, there is a user-generated video clip showing that a little girl is scared when she is watching a horror movie. The *scared* can be considered as an emotional event, which simultaneously causes a fearful facial expression in visual data, and a crying voice in audio data. Therefore, we use Sym-cHDP model incorporated with a symmetric correspondence mechanism to mine the latent emotional events between visual words and audio words within each video, which is very meaningful and helpful for video emotion analysis. The detailed modeling process will be specified in Sec. 4.

Note that Sym-cHDP is an unsupervised model, which means a supervised classifier should be joined up after this unsupervised learning process, for predicting a real emotional label for each video.

3. DESCRIPTION OF FEATURES

3.1. CNN-based Deep Features

The latest generation of CNNs, such as AlexNet [6], VGG [7] and GoogleLeNet [8], 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.

The latest study [11] experimentally demonstrates AlexNet and VGG deep architectures are more suitable for video emotion recognition than GoogleLeNet deep architecture. Besides, the features extracted from fully connected layers work more effectively than the ones extracted from convolutional layers. In our method, we therefore respectively draw the features from fully connected layers in AlexNet, VGG-F [7] and VGG-S [7] as our deep features to conduct the video emotion recognition tasks. Note that VGG model evolving from AlexNet model has several different architecture designs due to considering different accuracy/speed trade-offs. Here, two typical VGG models, *i.e.*, VGG-F and VGG-S, are selected for our work, where VGG-F focuses on speed, and VGG-S focuses on accuracy. Since there are two fully connected layers (*fc6* and *fc7*) existing in each of three CNN models, we totally evaluate six different combination patterns of deep features during the experimental phase.

3.2. Extraction of Visual Words and Audio Words

Since extracted deep features and MFCC features cannot be directly utilized in Sym-cHDP, we need to convert these features into independent visual words and audio words.

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 $\boldsymbol{D}_{fj} = (d_{fj1}, d_{fj2}, ..., d_{fjm})$, note that f is the index of video, and j is the index of image. We assume that each element d maps to an exclusive type of visual word, which means the vocabulary size of visual word V_D is equal to m. After *drop-out* processing, each element in deep feature vector becomes a nonnegative continuous variable, which can be considered as a likelihood of corresponding visual word. Then, we can draw the visual word count vector $\boldsymbol{n}_{fj}$ from a multinomial distribution:

$$\boldsymbol{n}_{fj} = (n_{fj1}, n_{fj2}, ..., n_{fjV_D}) \sim Multinomial(N, \boldsymbol{P}_{\boldsymbol{D}_{fj}}) \tag{1}$$

where N is the total number of visual words, and $N = \sum_{v=1}^{V_D} n_{fjv}$. $\boldsymbol{P}_{\boldsymbol{D}_{fj}}$ is a normalized representation for $\boldsymbol{D}_{fj}$. In this way, the visual word set X_{fj} for image j in video f can be filled with sampled N visual words. On video level, we can obtain the visual word set X_f for video f with $X_f = \{X_{f1}, X_{f2}, ..., X_{fJ_f}\}$, where J_f is the total image counts in video f.

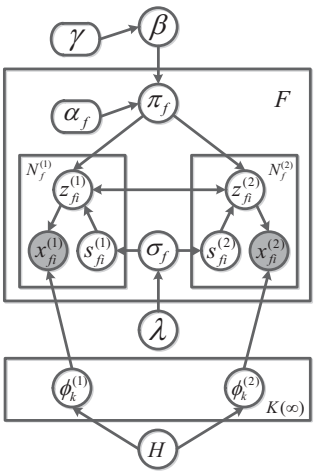

Fig. 2. Graphical representation for SBC of Sym-cHDP.

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

4. MULTIMODAL DATA MODELING

4.1. Generative Process

As we mentioned in Sec. 2, the video emotion can be expressed with a collection of latent emotional events occurring in the video, while each emotional event reflects in corresponding visual words and audio words. To cope with such multimodal data modeling issue for video emotion analysis, we therefore adopt a Sym-cHDP model [12], which incorporates a flexible symmetric correspondence mechanism [14] for modeling the generative process of video multimodal data. In Fig. 2, we illustrate a stick-breaking construction (SBC) [13] of Sym-cHDP with a graphical representation. Similar to original HDP, Sym-cHDP also has a two-layer hierarchy with global measure G and local measure G_f, where f denotes the index of the video. For SBC, we respectively use component (topic/emotional event) weight vectors $\boldsymbol{\beta}$ and $\boldsymbol{\pi_f}$ to construct corresponding measures in Sym-cHDP, the process is described as below:

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \quad G_f = \sum_{k=1}^{\infty} \pi_{fk} \delta_{\phi_k} \tag{2}$$

where each component weight vector consists of an infinite number of corresponding component weights, *i.e.*, $\boldsymbol{\beta} =$

$\{\beta_1, \beta_2, ..., \beta_k, ...\}$ and $\boldsymbol{\pi_f} = \{\pi_{f1}, \pi_{f2}, ..., \pi_{fk}, ...\}$. δ_{ϕ_k} is a probability measure concentrated at component ϕ_k.

Then, the generative process of Sym-cHDP is described with following steps:

Step 1: We draw the global component weight vector $\boldsymbol{\beta}$ from $GEM(\gamma)$ with a hyperparameter γ.

Step 2: Conditioned on $\boldsymbol{\beta}$ and hyperparameter α_f, each local component weight vector $\boldsymbol{\pi_f}$ is drawn from $DP(\alpha_f, \boldsymbol{\beta})$.

Step 3: For each video, we draw a multinomial pivot flag generator $\boldsymbol{\sigma_f}$ from $Dir(\lambda)$ with a hyperparameter λ.

Step 4: For the ith word of data type[1] l in video f, a pivot flag $s_{fi}^{(l)}$ is drawn from $Multinomial(\boldsymbol{\sigma_f})$.

Step 5: If $s_{fi}^{(l)} = l$, draw an emotional event $z_{fi}^{(l)}$ from $\boldsymbol{\pi_f}$. Otherwise if $s_{fi}^{(l)} = h \neq l$, draw an emotional event $z_{fi}^{(l)}$ from $Uniform(z_{f1}^{(h)}, ..., z_{fH_f^{(h)}}^{(h)})$.

Step 6: Conditioned on sampled $z_{fi}^{(l)}$, a word $x_{fi}^{(l)}$ is drawn from $f(x_{fi}^{(l)}|\phi_k^{(l)}, k = z_{fi}^{(l)})$.

Here, $GEM()$ indicates a GEM process [13], which is formed by such a process: $\hat{\beta}_k \sim Beta(1, \gamma)$, $\beta_k = \hat{\beta}_k \Pi_{i=1}^{k-1}(1 - \hat{\beta}_i)$. $DP()$ and $Dir()$ represent Dirichlet process and Dirichlet distribution, respectively. Different from other correspondence mechanisms [15], Sym-cHDP utilizes a multinomial pivot flag generator, which samples a pivot flag to control the way of generating the emotional event for the current word. For the word $x_{fi}^{(l)}$ with data type l, if its sampled pivot flag points at the same data type (*i.e.*, $s_{fi}^{(l)} = l$), the process of its emotional event assignment will be independent from other emotional event assignments and only associated with $\boldsymbol{\pi_f}$. However, if its sampled pivot flag points at the other data type h (*i.e.*, $s_{fi}^{(l)} = h \neq l$), its emotional event will be drawn from a uniform distribution that includes all the emotional events assigned to $H_f^{(h)}$ words with those pivot flags. Different from word counts $N_f^{(h)}$, $H_f^{(h)}$ is the number of words that have been already assigned with emotional events at the current step. Hence, $H_f^{(h)} \leq N_f^{(h)}$. Additionally, $f(x_{fi}^{(l)}|\phi_k^{(l)})$ indicates a word distribution conditioned on $\phi_k^{(l)}$. In fact, all the $\{\phi_k^{(l)}\}_{l=1}^2$ drawn from a base measure H share the same K emotional events.

Through such symmetric correspondence mechanism, Sym-cHDP can model the multimodal data with latent connections more flexibly, which is very beneficial to video emotion analysis.

4.2. Inference Method

In this section, we derive an inference method to estimate the latent emotional events and other variables within Sym-cHDP based on posterior representation sampler [16].

[1] In our work, we only consider visual and audio words, so $l \in \{1, 2\}$.

First of all, two component weight vectors $\boldsymbol{\beta}$ and $\boldsymbol{\pi_f}$ are sampled by:

$$\boldsymbol{\beta} = (\beta_1, ..., \beta_K, \beta_u) \sim Dir(T_{\cdot 1}, ..., T_{\cdot K}, \gamma) \tag{3}$$

$$\boldsymbol{\pi_f} = (\pi_{f1}, ..., \pi_{fK}, \pi_{fu}) \sim Dir(\tilde{\pi}_{f1}, ..., \tilde{\pi}_{fK}, \alpha_f \beta_u) \tag{4}$$

where every original infinite component weight vector is reformulated with a new augmentable finite vector that consists of K components and a promising component u. $T_{\cdot k}$ is a new variable that denotes the table counts in the Chinese restaurant franchise (CRF) representation [13] of Sym-cHDP, we can find more detailed explanation and sampling approach for $T_{\cdot k}$ in paper [13]. In addition, $\tilde{\pi}_{fk} = \alpha_f \beta_k + \sum_{(l)} C_{flk}^{(l)}$, where $C_{flk}^{(l)}$ denotes the counts for the word with data type l in video f possessing the pivot flag with the same data type l when the sampled emotional event is k.

As we described in Sec. 4.1, the pivot assignment affects the way of sampling emotional event in Sym-cHDP. We therefore derive a full conditional joint likelihood for estimating both emotional events $z_{fi}^{(l)}$ and pivot flag $s_{fi}^{(l)}$ for each word:

$$P(z_{fi}^{(l)} = k, s_{fi}^{(l)} = l | x_{fi}^{(l)}) \propto P(s_{fi}^{(l)} = l) P(z_{fi}^{(l)} = k) P(x_{fi}^{(l)} | z_{fi}^{(l)} = k)$$

$$= \begin{cases} \frac{C_{fl}^{-fli} + \lambda}{C_{fl}^{-fli} + \sum_{l' \neq l} C_{fl'} + 2\lambda} \cdot \pi_{fk} \cdot f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) & \text{if } k \text{ is used} \\ \frac{C_{fl}^{-fli} + \lambda}{C_{fl}^{-fli} + \sum_{l' \neq l} C_{fl'} + 2\lambda} \cdot \pi_{fu} \cdot f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) & \text{if } k \text{ is newborn} \end{cases} \tag{5}$$

$$P(z_{fi}^{(l)} = k, s_{fi}^{(l)} = h | x_{fi}^{(l)}) \propto P(s_{fi}^{(l)} = h) P(z_{fi}^{(l)} = k) P(x_{fi}^{(l)} | z_{fi}^{(l)} = k)$$

$$= \frac{C_{fh}^{-fli} + \lambda}{C_{fh}^{-fli} + \sum_{h' \neq h} C_{fh'} + 2\lambda} \cdot \frac{n_{fk}^{(h)}}{N_f^{(h)}} \cdot f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) \tag{6}$$

where C_{fl} denotes the counts for the pivot flag pointing at data type l over all the multimodal data in video f, and a superscript $-fli$ appearing in C_{fl} (*i.e.*, C_{fl}^{-fli}) means the removal of the pivot flag count for the word i. $n_{fk}^{(h)}$ denotes the counts for emotional event k assigned to words of data type h in video f. $f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ and $f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ are two different types of conditional word likelihood functions conditioned on emotional event k. When the k is previously used, $f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ is formulated with:

$$f_k^{-x_{fi}^{(l)}}(x_{fi}^{(l)} = v^{(l)}) = \int f(x_{fi}^{(l)} = v^{(l)} | \phi_k^{(l)}) p(\phi_k^{(l)} | X_{lk}^{-fli}, H) \, d\phi_k^{(l)}$$

$$= \frac{n_{kv^{(l)}}^{-fli} + \tau}{\sum_{v^{(l)}} n_{kv^{(l)}}^{-fli} + V^{(l)}\tau} \tag{7}$$

when the k is newborn, $f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)})$ is formulated with $f_{k^{new}}^{-x_{fi}^{(l)}}(x_{fi}^{(l)}) = \frac{1}{V^{(l)}}$, 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 $x_{fi}^{(l)}$. $n_{kv^{(l)}}^{-fli}$ is the counts for the word $v^{(l)}$ assigned to emotional event k

Table 1. Unimodel Results with Different Deep Features

Methods		CNN	HDP(0.1)	HDP(0.5)	HDP(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

except for $x_{fi}^{(l)}$. $V^{(l)}$ is the vocabulary size of data type l. τ is a hyperparameter for a Dirichlet distribution, and $H = Dir(\tau)$.

Finally, we apply Gibbs sampling approach to implement this inference method for Sym-cHDP. In this way, the latent emotional events can be well learned from observed visual words and audio words within each video.

5. EXPERIMENTS

5.1. Experimental Setup

In this experiment, an *Acted Facial Expressions in the Wild* (AFEW)[2] dataset that comprises 957 video clips extracted from movies, is adopted for evaluation. In AFEW dataset, several key frames deemed as image data are drawn from each video clip for emotion recognition. Videos are categorized with seven basic emotions containing *angry*, *happy*, *disgust*, *fear*, *sad*, *surprise* and *neutral*.

For deep feature extraction, we adjust the dimension m of *fc6* layer and *fc7* layer for each CNN from 4096 to 1024 to cut the computation cost, which means the vocabulary size of visual word V_D is specified to 1024. For each key frame, we set the total number of extracted visual words N to 5120. For audio MFCC feature extraction, the vocabulary size of audio word V_M is set to 1000.

For the initialization of nonparametric topic models (*i.e.*, HDP, Sym-cHDP and other baseline models), the hyperparameters γ and α_f are sampled from a gamma prior $Gamma(1.0, 1.0)$, and updated every iteration. The Gibbs sampling system will totally run 1000 iterations so as to let all the variables in the model fully converge.

A simple LIBLINEAR[3] classifier is finally utilized in our framework for video emotion classification. We conduct evaluations with a five-fold cross-validation scheme. The performance is measured by average classification accuracy.

5.2. Unimodal Analysis on Deep Features

In this section, we focus on unimodal (only image data used) emotion analysis by using CNN and HDP, to evaluate the performance on different CNN models and different types of deep features within these CNN models, as shown in Tab. 1.

[2]https://cs.anu.edu.au/few/AFEW

[3]https://www.csie.ntu.edu.tw/ cjlin/liblinear/

Table 2. 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)	25.52	26.53	29.94	28.21	34.34	32.91
Corr-HDP(Audio)	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**

Evaluation on Deep Features with HDP. HDP as a unimodel version of Sym-cHDP is utilized for evaluating the emotion classification performance on different types of deep features. In this experiment, we respectively assign the controlling parameter τ with $\tau = 0.1$, $\tau = 0.5$ and $\tau = 1.0$, to conduct three different sets of experiments. According to Tab. 1, for the cases of using VGG-F and VGG-S, the deep features drawn from the *fc6* layer significantly outperform the ones drawn from the *fc7* layer, while such phenomenon is not so clear in the case of using AlexNet. This indicates that the *fc6* layer may involve more sentimental information than the *fc7* layer for the same image. Besides, we can find that the HDP performs the best when $\tau = 0.5$.

Comparison between CNN methods and HDP methods. To validate the effectiveness of HDP model, we also conduct three comparative experiments by directly using three different CNN models. The results demonstrate that the optimized HDP outperforms CNN method by 4.47%, 2.96% and 6.01%, when using AlexNet, VGG-F and VGG-S, respectively. This indicates that the learning process for latent emotional events in HDP can actually boost the emotion recognition performance. Besides, we observe that VGG-S performs the best among all the CNN models, so we infer that the VGG-S architecture is more suitable for the emotion recognition task.

5.3. Multimodal Analysis

Finally, comprehensive experiments based on video multimodal data (visual data and audio data) are conducted for evaluating our method and other baselines. Here, we select two styles of baseline sets: the baseline models in the first set are general classifiers including LIBLINEAR and SVM, where the deep features and MFCC features are simply fused without being further learned, the baseline models in the second set are other nonparametric topic models including conditionally independent HDP (CI-HDP) inspired by paper [17], and correspondence HDP (Corr-HDP) inspired by paper [15]. In the initialization phase, we set $\tau = 0.5$ to all the nonparametric topic models, and set $\lambda = 1.0$ to Sym-cHDP. The experimental results are shown in Tab. 2.

Compared with both LIBLINEAR and SVM, our method with Sym-cHDP significantly outperforms those general classifiers for every scenario, which demonstrates that the mined latent emotional events from visual words and audio word-

s via Sym-cHDP are very useful for video emotion recognition. On the other hand, Sym-cHDP also outperforms other nonparametric topic models CI-HDP and Corr-HDP for every scenario, which demonstrates that the flexible symmetric correspondence mechanism can make Sym-cHDP work more effectively on multimodal data modeling.

In addition, we also find that Sym-cHDP working on multimodal data significantly outperforms HDP working on unimodel data (visual data). This shows that the audio MFCC features are very complementary to the deep features for such video emotion recognition task, since they record the sentiment from different angles.

6. CONCLUSIONS

This paper presents a novel Bayesian nonparametric multimodal data modeling framework to learn the emotions from videos. In this framework, we first draw the CNN-based deep features and the audio MFCC features from each video. Then, a symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) is utilized to model the multimodal data, and furthermore learn the latent emotional events between image data and audio data. We finally demonstrate that our method outperforms other baselines via comprehensive experimentations. Future work may focus on applying a supervised nonparametric topic modeling approach, which can directly estimate a emotion label without using an extra classifier.

7. REFERENCES

[1] Stacy C Marsella and Jonathan Gratch, "Ema: A process model of appraisal dynamics," *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, 2009.

[2] Hang-Bong Kang, "Affective content detection using hmms," in *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003, pp. 259–262.

[3] Min Xu, Changsheng Xu, Xiangjian He, Jesse S Jin, Suhuai Luo, and Yong Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2013.

[4] René Marcelino Abritta Teixeira, Toshihiko Yamasaki, and Kiyoharu 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.

[5] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue, "Predicting emotions in user-generated videos.," in *AAAI*, 2014, pp. 73–79.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[7] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proceedings of the British Machine Vision Conference*, 2014.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[9] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *CoRR*, 2014.

[10] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 381–388.

[11] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 15–22.

[12] Koji Eguchi, Kosuke Fukumasu, Tsukasa Omoto, and Eric P Xing, "Symmetric correspondence topic models for multimodal data analysis," *ArXiv*, 2017.

[13] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei, "Hierarchical Dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, 2006.

[14] Kosuke Fukumasu, Koji Eguchi, and Eric P Xing, "Symmetric correspondence topic models for multilingual text analysis," in *Advances in Neural Information Processing Systems*, 2012, pp. 1286–1294.

[15] Jianfei Xue and Koji Eguchi, "Sequential correspondence hierarchical Dirichlet processes for video data analysis," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 229–233.

[16] Yee Whye Teh, "Dirichlet process," in *Encyclopedia of machine learning*, pp. 280–287. Springer, 2010.

[17] Elena Erosheva, Stephen Fienberg, and John Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5220–5227, 2004.