PERSONALIZED VIDEO EMOTION TAGGING THROUGH A TOPIC MODEL

Shan Wu⋆ *Shangfei Wang*⋆† *Zhen Gao*⋆

⋆ Key Lab of Computing and Communication Software of Anhui Province,
School of Computer Science and Technology,
University of Science and Technology of China
Hefei, Anhui, China
Email: sa14ws@mail.ustc.edu.cn, sfwang@ustc.edu.cn, gzgqllxh@mail.ustc.edu.cn

ABSTRACT

The inherent dependencies among video content, personal characteristics, and perceptual emotion are crucial for personalized video emotion tagging, but have not been thoroughly exploited. To address this, we propose a novel topic model to capture such inherent dependencies. We assume that there are several potential human factors, or "topics," that affect the personal characteristics and the personalized emotion responses to videos. During training, the proposed topic model exploits the latent space to model the relationships among personal characteristics, video content and video tagging behaviors. After learning, the proposed model can generate meaningful latent topics, which help personalized video emotion tagging. Efficient learning and inference algorithms of the model are proposed. Experimental results on the CP-QAE-I database demonstrate the effectiveness of the proposed approach in modeling complex relationships among video content, personal characteristics, and perceptual emotion, as well as its good performance in personalized video emotion.

Index Terms— Video emotion tagging; personal characteristics; topic model

1. INTRODUCTION

Recent years have seen a rapid increase in the size of video collections due to the popularity of the Internet and of portable cameras, such as the smart-phone. These video collections have become the medium for many people to communicate and to find entertainment with the development of online services like YouTube and Vimeo. Since emotion is the key factor during communication and entertainment, video emotion tagging has begun to attract more attention.

Current video emotion tagging can be divided into two approaches: direct and implicit. Direct approaches sign the emotion labels to videos directly from related audiovisual features, while implicit approaches infer emotion labels of videos based on an automatic analysis of a user's spontaneous response when watching the videos. The tagged emotions can be either expected emotions or induced emotions. Expected emotions can be regarded as common emotions, which are communicated through visual and aural elements based on film grammar. Induced emotions, however, are the audiences' emotions elicited during watching videos. They are personalized emotions, since the same video may induce different emotions from different audiences due to their various personalities and culture backgrounds. A comprehensive survey on video emotion tagging can be found in [1].

Compared with direct approaches, implicit video tagging is more personal and subjective, since users' spontaneous nonverbal responses when watching the videos provide clues of actual emotions induced by the videos. However, it is inconvenient to collect users' physiological signals during emotion tagging due to the high cost of physiological sensors and the discomfort of users. The mainstream research on direct video emotion tagging assigns common emotion tags to videos based on video content. Little research performs personalized video emotion tagging. Yoo and Cho [2] propose to use an interactive genetic algorithm for an individual video scene retrieval. By integrating users' evaluation into this genetic algorithm, their method can retrieve videos which satisfy the user's individual emotion query. Zhang et al. [3] incorporate a user's feedback, profile, and affective preference to realize an integrated system for personalized music video affective analysis. A downside of these two works is that they both require users' feedback to realize personalized video affective analysis, which increases users' burden. Wang et al. [4] adopt the Bayesian network to model the relationship between personalized emotion tags and video's common emotion tags to predict the video's personalized tags. Their work omits personal characteristics data, which is crucial for personalized emotion tagging.

Since emotion is a person's subjective evaluation of a stimulus event (in this case, a video), personalized video emotion tagging should involve video content, personal characteristics data and personalized emotion labels. As yet, no research has fully explored the relations among the three. The lack of such research is caused by both the difficulty of this task and the lack of an appropriate database. Only re-

† is the corresponding author.

cently, Guntuku et al. [5] constructed the CP-QAE-I database, including videos, personal characteristics data, and personalized emotion labels. The CP-QAE-I database contains 144 video sequences based on 12 short movie clips varying by frame rate, frame dimension, and bit rate. There are 114 participants. Every participant gives 5-point ratings to each watched video sequences as personalized video tags. In total, 1232 records are collected. The database also provides 11 personal characteristics of each participant, including six cultural traits and five personality traits. For details, please refer to [5]. After the release of the CP-QAE-I database, Scott et al. [6] used this database to analyze the influence of personality and cultural traits on the perception of video quality and subsequent enjoyment. Their analysis verifies the key role of personal characteristics in the perception of video quality and enjoyment. But this work does not perform personalized video emotion tagging on the CP-QAE-I database.

Therefore, in this paper, we propose a new topic model to capture the probabilistic dependencies among video content, personal characteristics data, and perceptual emotion. Specifically, we suppose that there are potential human factors that influence the personalized video emotion response. These potential human factors are "topics" in our model, which exploits the latent space to capture the relationships between certain personal characteristics and personalized video emotion labels. Efficient learning and inference algorithms are proposed. Experimental results on the CP-QAE-I database demonstrate that our model can generate meaningful topics, and improve the performance of personalized video emotion tagging.

Compared to related work, we are the first to capture the dependencies among video content, personal characteristics, and perceptual emotions for personalized video emotion tagging. Furthermore, we propose a new topic model for personalized video emotion tagging as well as efficient learning and inference algorithms of our model.

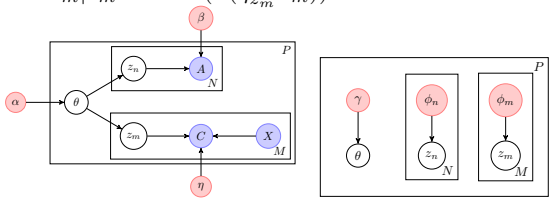
2. PERSONALIZED VIDEO TAGGING

2.1. Proposed topic model

We suppose that there are some potential human factors, which influence personal characteristics and personalized feelings about videos. We take these potential factors as "topics". We leverage the topic model using the latent space to model these potential factors, capturing the relationships between the personal characteristics and personalized video tagging behaviors.

Our model is shown in Figure 1(a). There are N personal characteristics and M video tagging records for each person. Each video is tagged with the label $C \in \{0, 1\}$, where 0 indicates low level and 1 indicates high level. The number of topics is K, which is fixed. $\alpha_{1:K}$ is the parameters of the Dirichlet distribution for K topics. A represents the personal characteristics, and a_n is a one-hot vector to indicate the state of the nth characteristic. θ, z_n, z_m are the latent variables. The variable z is the one-hot vector to indicate the topic assignment. The probability of the nth characteristic is parameterized by a $k \times S^n$ matrix β^n, where S^n is the number of the states of the nth characteristic, and $\beta^n_{ij} = p(a_{ni} = 1 | z_{nj} = 1)$. X represents the video features and $\eta_{1:K}$ is coefficients for different topics. The detailed generating process is as follows:

1. Draw topic proportions $\theta \sim Dir(\alpha)$.
2. For each personal characteristic $a_n, n \in \{1, 2, ..., N\}$:
   (a) Draw the topic assignment $z_n | \theta \sim Mult(\theta)$.
   (b) Draw the personal characteristic $a_n | z_n \sim Mult(\beta^n_{z_n})$.
3. For each movie tagged by this person $c_m, x_m, m \in \{1, 2, ..., M\}$:
   (a) Draw the topic assigment $z_m | \theta \sim Mult(\theta)$.
   (b) Draw the personalized video emotion tags $c_m | z_m \sim Beta(\sigma(\eta^T_{z_m} x_m))$.



(a) (b)

Fig. 1. (a) proposed topic model. (b) variational distribution.

Given the parameters α, β, η and the video features X, the joint probability of the mixture "topic" proportions θ, M video tags C, N personal characteristics A, and the topic assignments Z is shown as follows:

$$p(\theta, Z, A, C | \alpha, \beta, \eta, X) = p(\theta|\alpha) \prod_{n=1}^{N} \{p(z_n|\theta) \cdot p(a_n|z_n, \beta^n)\} \prod_{m=1}^{M} \{p(z_m|\theta) \cdot p(c_m|z_m, \eta, x_m)\} \quad (1)$$

where

$$p(\theta|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} \cdot \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \quad (2)$$

$$p(z_n|\theta) = \prod_{i=1}^{k} \theta_i^{z_{ni}} \quad (3)$$

$$p(a_n|z_n, \beta) = \prod_{i=1}^{k} \prod_{j=1}^{S^n} (\beta^n_{ij})^{z_{ni} \cdot a_{nj}} \quad (4)$$

$$p(z_m|\theta) = \prod_{i=1}^{k} \theta_i^{z_{mi}} \quad (5)$$

$$p(c_m|z_m, \eta, x_m) = \prod_{i=1}^{k} [\sigma(\eta_i^T x_m)^{c_m} (1 - \sigma(\eta_i^T x_m))^{1 - c_m}]^{z_{mi}} \quad (6)$$

Compared to LDA [7], our model has two major differences. First, our personal characteristic variable A is different from the "words" in the topic model. In LDA, different words for a topic are generated from one multinomial distributions. In our model, each personal characteristic has a multinomial distributions for one topic, as each person will be in one state of this characteristic and different personal characteristics may have different sizes of states. Second, compared to different variants of sLDA [8][9], our model constructs K classifiers based on the raw video features while sLDA constructs only one classifier based on the distributions of latent variable z as new features. Our model is like the mixture of experts model and the latent variable θ for each person has the same function as the mixture coefficients.

2.2. Model Learning

For inference and parameter estimation, we need to compute the conditional probability of observed samples given the parameters:

$$p(A, C|\alpha, \beta, \eta, X) = \int_\theta \sum_z p(\theta, z, A, C|\alpha, \beta, \eta, X) d\theta \tag{7}$$

As it is intractable for exact inference, we employ the variational inference like LDA[7].

2.2.1. Variational Inference

The graphical model representation of the variational distributions is shown in Figure 1(b), which is defined as follows:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \prod_{m=1}^{M} q(z_m|\phi_m) \tag{8}$$

where ϕ_n, ϕ_m are both multinomial distributions over K topics, ϕ_n is for the personal characteristics and ϕ_m is for the video tagging behaviors, and γ is the parameters for Dirichlet distributions.

Let $\Theta = \{\alpha, \beta, \eta\}$. Following the Jensen's inequality to bound the log likelihood, we obtain:

$$\begin{aligned} log\, p(A, C|\Theta) &= log \int_\theta \sum_z \frac{p(A, C, \theta, z|\Theta, X) q(\theta, z)}{q(\theta, z|\gamma, \phi)} d\theta \\ &\geq \int_\theta \sum_z q(\theta, z)\, log\, \frac{p(A, C, \theta, z|\Theta, X)}{q(\theta, z)} d\theta \\ &= E_q[log\, p(A, C, \theta, z|\Theta, X)] - E_q[log\, q(\theta, z)] \\ &= L(\Theta, \gamma, \phi) \end{aligned} \tag{9}$$

It can be proved that maximizing L is equivalent to minimizing the KL divergence between q and the true posterior probability. Similar to [7], the updated equations for γ and ϕ_n are shown as follows:

$$\phi_{ni} \propto exp(\Psi(\gamma_i) - \Psi(\gamma_0)) \beta^n_{ia_n} \tag{10}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} + \sum_{m=1}^{M} \phi_{mi} \tag{11}$$

Let the derivative of $L_{[\phi_{mi}]}$ with respect to ϕ_m equal zero, and add the constraints $\sum_{i=1}^{k} \phi_{mi} = 1$, then we obtain the updated equation for ϕ_m:

$$\phi_{mi} \propto exp(\Psi(\gamma_i) - \Psi(\gamma_0)) \sigma(\eta_i^T x)^{c_i} (1 - \sigma(\eta_i^T x))^{1 - c_i} \tag{12}$$

Given the personal characteristics and the video enjoyment records for one person, we use the variational inference to update the ϕ_n, ϕ_m and γ based on Equation 10, 12 and 11 respectively until L converges.

2.2.2. Estimating the parameters

Given the training data $D = \{(A_i, C_i)_{i=1}^P, X\}$, the log-likelihood function is defined as follows:

$$\ell(\alpha, \beta, \eta) = \sum_{i=1}^{P} log\, p(A_i, C_i|\alpha, \beta, \eta, X) \tag{13}$$

We use the variational EM to estimate the parameters. In the E-step, we apply the variational inference to find the approximate posterior distribution q for the latent variables. In the M-step, we find the maximum-likelihood estimation of the β and γ based on the posterior distribution q. We repeat this process until convergence. The hyper-parameters α is fixed as $\frac{10}{K}$ during training.

To maximize the log likelihood respect to β, we select the terms containing β and add the constraints:

$$\ell_{[\beta^n_{is}]} = \sum_{p=1}^{P} \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{s=0}^{S^n} \phi_{pni} a^s_{pn}\, log\, \beta^n_{is} + \sum_{n=1}^{N} \sum_{i=1}^{k} \lambda_{ni} (\sum_{s=0}^{S^n} \beta^n_{is} - 1) \tag{14}$$

Let the derivative with respect to β^n_{is} equal zero, we obtain:

$$\beta^n_{is} \propto \sum_{p=1}^{P} \phi_{pni} a^s_{pn} \tag{15}$$

For the $\eta_{1:K}$, the terms containing it are:

$$\ell_{[\eta]} = \sum_{p=1}^{P} \sum_{m=1}^{M} \sum_{i=1}^{k} \phi_{pmi} [c_{pm}\, log\, \sigma(\eta_i^T x) + (1 - c_{pm})\, log(1 - \sigma(\eta_i^T x))] \tag{16}$$

We use the LBFGS [10] to find the optimal η using the derivative defined as follows:

$$\frac{\partial L}{\partial \eta_i} = \sum_{p=1}^{P} \sum_{m=1}^{M} \phi_{pmi} [c_{pm} - \sigma(\eta_i^T x_m)] x_m \tag{17}$$

2.3. Model Inference

For the person who already appears in the training set, we can predict the personalized video emotion tags according to:

$$p(c = 1|P_i, x_{new}) = \sum_{j=1}^{K} p(c = 1|\eta_i, x_{new}) q_i(\theta_j|\gamma) \tag{18}$$

where P_i means the ith person and q_i is the approximate posterior distribution obtained for P_i during training.

For a new person with personal characteristics, we first use the variational inference to obtain the approximate posterior distribution. We iteratively update ϕ_n and γ with Equation 10 and Equation 19 respectively until convergence.

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{19}$$

After that, we predict the personalized emotion tags for this person with a new video sample using Equation 18.

3. EXPERIMENTS AND ANALYSIS

3.1. Experimental Conditions

As described in Section 1, the CP-QAE-I database is the only database providing video content, cultural and personality traits, and personalized video emotion tags. Therefore, we conduct experiments of personalized video emotion tagging on the CP-QAE-I database to validate our proposed model. We adopt the subjective enjoyment label as the personalized emotion tags. For simplicity, we divide each characteristics into two states based on the mean score. For personality traits, the two states of openness can be described as inventive and consistent. It is similar for conscientiousness(efficient or careless), extroversion(outgoing or solitary), agreeableness(friendly or analytical), and neuroticism(sensitive or confident) as defined in FFM[11]. Each cultural trait is divided into high and low states. For the personalized emotion tags, we take 1-2 point as "low" state while 3-5 points is considered "high" state.

Both audio and visual features are extracted from videos. Specifically, we extract commonly used audio features includ-

ing spectrum flux, Zero Crossing Rate (ZCR), average energy, average energy intensity, standard of deviation of ZCR and Mel-frequency Cepstral Coefficients (MFCCs). For visual features, we extract visual excitement, lighting key, and color energy[12], which have proved to be powerful at affecting the emotions of the viewers. The C-QAE-I database provides the frame-rate (FR), frame dimension (Dim) and bit rate (BR) for each video. These video quality parameters (FR, Dim, BR, FR * Dim, FR * BR, Dim * BR, FR * Dim * BR) are also used as video features.

3.2. Analysis on the latent topics

We adopt the leave-one-video-out cross-validation method to analyze the influence of the latent topic number on personalized video emotion tagging. The accuracy of personalized video emotion tagging varies with the number of topics. Specifically, the accuracy quickly increases from 68.1% to 71% when K varies from 1 to 3. This proves the effectiveness of the latent topics for personalized emotion prediction. When K changes from 5 to 12, the accuracy of our model varies from 71.5% to 72.5%. The sliding interval is quite small, demonstrating that when K is big enough, the proposed model is not very sensitive to the number of topics . According to John Holland's theory[13], most personalities of people is a combination of six basic personality types: realistic, investigative, artistic, social, enterprising, and conventional. Therefore, we set the number of topics to six for our experiments.

Table 1. Personality Traits for Each Topic

topic1	topic2	topic3	topic4	topic5	topic6
inventive	inventive	consistent	consistent	consistent	inventive
careless	efficient	efficient	careless	efficient	careless
solitary	solitary	outgoing	outgoing	solitary	outgoing
analytical	analytical	friendly	friendly	analytical	friendly
sensitive	confident	confident	sensitive	sensitive	sensitive

Like LDA, our model can generate the personal characteristic tags for each topic. Take personality traits for example, Table 1 shows the state of each personal trait with highest probability for each topic. From Table 1, we find that the personality descriptions for each topic are consistent with the John Holland's theory in some sense. Specifically, topic 1 matches the type of artistic, topic 2 matches investigative, topic 3 is social, topic 4 is realistic, topic 5 indicates conventional, and topic 6 may match enterprising. It means the proposed model generates meaningful latent topics from users' personal characteristics data, demonstrating the effectiveness of the proposed model in capturing complex relations among video content, personal characteristics, and perceptual enjoyment. This leads to good performance of personalized video enjoyment tagging in the following sections.

3.3. Personalized video emotion tagging

We conduct a leave-one-video-out cross-validation experiment and a leave-one-subject-out cross-validation experiment to evaluate personalized video emotion tagging performance of the proposed model for a existing person and a new person respectively. For the first experiment, we compare our model with a support vector machine (SVM) classifier, which uses both personal characteristics and video features as the input with a linear kernel. For the second experiment, in addition to comparing with SVM, we also compare with the method which predicts emotion labels from the video features by using the highest probability across K classifiers based on the $\eta_{1:K}$ in our model. We refer to this as method*, which does not use the personal characteristics. The results of the two personalized video emotion tagging experiments are shown in Table 2, and we can find that:

Firstly, compared to SVM, our method improves the accuracy, F1score, and Kappa by 4.4%, 0.023, and 0.074 respectively in the leave-one-video-out cross-validation experiment, as well as 2.79%, 0.034 and 0.057 in the leave-one-subject-out cross-validation experiment. Instead of directly using the personal characteristics as features like SVM, our model exploits the latent space to capture the structure of potential human factors. This leads to better performance.

Secondly, the performances of both SVM and the proposed model are worse for a new subject. It is reasonable that personalized video emotion tagging for a new subject is more challenging, since there is no information on the subject in the training set.

Finally, in leave-one-subject-out experiment, Method*, which ignores user difference, performs the worst. It predicts the same video sample as the same label across different viewers, which is unreasonable.

Table 2. Experimental results of personalized video emotion tagging.

	leave-one-video-out				leave-one-subject-out					
	SVM		Our method		SVM		Method*		Our method	
	Low	High	Low	High	Low	High	Low	High	Low	High
Low	322	189	308	203	296	215	227	284	314	197
High	205	516	139	582	217	504	163	558	201	520
Acc.	68.0%		72.2%		64.9%		63.7%		67.69%	
F1.	0.620		0.643		0.578		0.504		0.612	
Kappa	0.344		0.418		0.278		0.226		0.335	

4. CONCLUSION

The inherent relationships among video content, personal characteristics, and personalized video emotion tags are crucial for personalized video emotion. In this work, we propose a new topic model capturing these inherent relationships. On the training phase, our model exploits the latent space modeling relationships among video content, personal characteristics, and perceptual emotion. On the testing phase, our model can infer subjective video emotion tags from video content and personal characteristics. The results on the CP-QAE-I database demonstrate that our model can generate meaningful latent topics, and improve the performance of personalized video enjoyment tagging.

ACKNOWLEDGEMENT

This work has been supported by the National Science Foundation of China (Grant No. 61473270, 61228304, 61175037), and the project from Anhui Science and Technology Agency (1508085SMF223).

5. REFERENCES

[1] Shangfei Wang and Qiang Ji, "Video affective content analysis: a survey of state of the art methods," *Affective Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[2] Hun-Woo Yoo and Sung-Bae Cho, "Video scene retrieval with interactive genetic algorithm," *Multimedia Tools and Applications*, vol. 34, no. 3, pp. 317–336, 2007.

[3] Shiliang Zhang, Qingming Huang, Shuqiang Jiang, Wen Gao, and Qi Tian, "Affective visualization and retrieval for music video," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 510–522, 2010.

[4] Shangfei Wang, Zhilei Liu, Yachen Zhu, Menghua He, Xiaoping Chen, and Qiang Ji, "Implicit video emotion tagging from audiences facial expression," *Multimedia Tools and Applications*, pp. 1–28, 2014.

[5] Sharath Chandra Guntuku, Michael James Scott, Huan Yang, Gheorghita Ghinea, and Weisi Lin, "The cp-qae-i: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia," Costa Navarino, Messinia, Greece, 2015.

[6] Michael James Scott, Sharath Chandra Guntuku, Yang Huan, Weisi Lin, and Gheorghita Ghinea, "Modelling human factors in perceptual multimedia quality: On the role of personality and culture," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, 2015, pp. 481–490.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[8] Jon D. Mcauliffe and David M. Blei, "Supervised topic models," in *Advances in Neural Information Processing Systems 20*, J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, Eds., pp. 121–128. Curran Associates, Inc., 2008.

[9] Chong Wang, David Blei, and Fei-Fei Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1903–1910.

[10] José Luis Morales and Jorge Nocedal, "Remark on algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound constrained optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 38, no. 1, pp. 7, 2011.

[11] Lewis R Goldberg, "An alternative" description of personality": the big-five factor structure.," *Journal of personality and social psychology*, vol. 59, no. 6, pp. 1216, 1990.

[12] Hee Lin Wang and Loong-Fah Cheong, "Affective understanding in film," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 6, pp. 689–704, 2006.

[13] John L Holland, *Making vocational choices: A theory of vocational personalities and work environments* ., Psychological Assessment Resources, 1997.