Inferring Emotional Tags From Social Images With User Demographics

Boya Wu, *Student Member, IEEE,* Jia Jia, *Member, IEEE,* Yang Yang, Peijun Zhao, Jie Tang, *Senior Member, IEEE,* and Qi Tian, *Fellow, IEEE*

***Abstract*—Social images, which are images uploaded and shared on social networks, are used to express users' emotions. Inferring emotional tags from social images is of great importance: it can benefit many applications, such as image retrieval and recommendation. Whereas previous related research has primarily focused on exploring image visual features, we aim to address this problem by studying whether user demographics make a difference regarding users' emotional tags of social images. We first consider how to model the emotions of social images. Then, we investigate how user demographics, such as gender, marital status and occupation, are related to the emotional tags of social images. A partially labeled factor graph model named the demographics factor graph model (D-FGM) is proposed to leverage the uncovered patterns. Experiments on a data set collected from the world's largest image sharing website Flickr[1] confirm the accuracy of the proposed model. We also find some interesting phenomena. For example, men and women have different patterns to tag "anger" for social images.**

***Index Terms*—Emotion, image, user demographics.**

I. Introduction

Emotion plays a major role in our daily life. It stimulates the mind 3,000 times faster than rational thought and influences our decisions [1]. With the rapid development of image social networks such as Flickr[1] and Instagram[2], people are becoming used to sharing their emotional experiences through images on these platforms. Our preliminary statistics indicate that 38% of the image tags written by the owners of images on the world's largest image social network Flickr contain either positive or negative emotional words. As shown in Fig. 1, an image depicting a heavy rainstorm may be tagged as sadness, whereas an image showing colorful balloons may be tagged as happiness. We define images that are uploaded and shared on social networks as "**social images**".

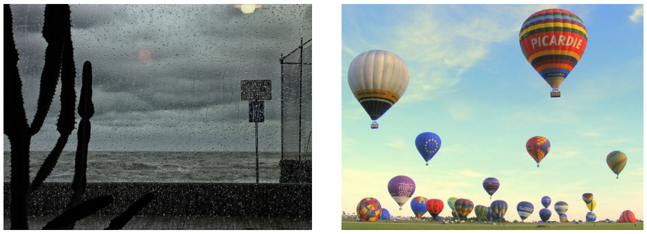


Fig. 1. An image depicting a heavy rainstorm may express a low mood, whereas an image showing colorful balloons may express happiness.

Inferring emotional tags for social images will benefit a number of applications, such as image retrieval based on their emotional contents, tag suggestion and image annotation [2], [3], [4], [5].

To date, considerable research effort has been devoted to inferring emotions from various types of inputs, including images [6], [7], [8], [9], [10], [11], [12], speech [13], [14], [15], [16], and audio-visual data [17], [18], [19], [20], [21], [22], [23], [24], [25]. These studies focus on modeling emotions, extracting effective features and utilizing diverse types of learning methods.

Moreover, recent research on social networks has verified that the demographics of users are associated with the behaviors of users. Dong *et al.* [26] reported that people of different ages have different social strategies for maintaining their social connections. Huang *et al.* [27] uncovered how user demographics influence the formation of closed triads on social networks. Moreover, psychological and behavioral research has proven that human perception of emotions varies according to their personal attributes. Fischer *et al.* [28] noted that there is a gender difference in the perception of emotions, namely, men report more powerful emotions (*e.g.*, anger), whereas women report more powerless emotions (*e.g.*, sadness and fear). Simon and Nath [29] found that girls talked more about the emotional aspects of their experiences than boys did in early parentchild conversations. Similar works can be found in [30], [31], [32]. These findings stimulate our curiosity. Will it be possible to utilize user demographics to improve the accuracy of inferring emotional tags from social images?

The problem is non-trivial and presents us with several challenges. First, although a few works demonstrate the existence of a correlation between the demographics and emotions of users, it is still unclear whether the correlation exists

Manuscript received February 29, 2016; accepted January 04, 2017. This work is supported by National Key Research and Development Plan (2016YFB1001200), the Innovation Method Fund of China (2016IM010200). This work is supported by the National Natural Science Foundation of China (61370023, 61521002, 61561130160, 61602033, 61561130160). This work is partially supported by the major project of the National Social Science Foundation of China(13&ZD189).

Boya Wu, Jia Jia and Peijun Zhao are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, Key Laboratory of Pervasive Computing, Ministry of Education, Tsinghua National Laboratory for Information Science and Technology (TNList) (e-mail: wuboya10@gmail.com, jjia@tsinghua.edu.cn, 421769833@qq.com).

Yang Yang is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. (email: yangya@zju.edu.cn))

Jie Tang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (email: jietang@tsinghua.edu.cn)

Qi Tian is with the Department of Computer Science, the University of Texas at San Antonio (UTSA) (e-mail: qitian@cs.utsa.edu)

[1] http://www.flickr.com/

[2] http://www.instagram.com

on image social networks. Second, how can we model user demographics and other information (visual features, social correlations, and so forth) in a joint framework? Third, how can we validate the effectiveness of the proposed model on a real-world image social network?

To address the above challenges, we randomly download 2,060,353 images and 1,255,478 users from the world's largest image sharing website Flickr. With the constraint that images must be associated with necessary information, we construct a data set containing 854,734 images uploaded by 3,181 users. On average, a user uploads 269 images. We first consider how to model the emotions of social images and unveil nine major emotion categories, namely, awe, amusement, contentment, excitement, anger, disgust, sadness, fear and boredom. Then, we investigate whether user demographics such as gender, marital status and occupation are related to the emotional tags of social images. We uncover several patterns, and a partially labeled factor graph model named the demographics factor graph model (**D-FGM**) is proposed to leverage user demographics in the modeling as different factors. The experimental results confirm the accuracy of the proposed model (0.2984), achieving a +0.126 improvement compared with naive Bayesian (0.1724) and a +0.1382 improvement compared with SVM (support vector machine, 0.1602). The effectiveness of the user demographics factors is also demonstrated by the factor contribution analysis, which reveals some interesting behavioral phenomena. For example, in terms of *amusement*, the emotional tags are primarily determined by visual features, indicating that although users may be of different gender, marital status and occupation, they tend to have similar patterns to tag amusement. Interestingly, however, when considering *anger*, males and females have different tagging patterns, which corresponds to the findings reported in behavioral research [28], [29]. Regarding *fear* and *sadness*, whether users are single or married influences the emotional tags. For *contentment* and *boredom*, the emotional tags are associated with the users' occupations.

The remainder of this paper is organized as follows. In Section II, we survey the existing research in the area of image emotion inference and user demographics. In Section III, we formally formulate the problem. In Section IV, we introduce the emotional social image data that we establish. In Section V, we verify the correlation between user demographics and emotional tags of social images and in modeling and provide an overview of the proposed D-FGM. In Section VII, we conduct experiments and report the experimental results. In Section VIII, we conclude this work and discuss ideas for future work.

II. Related works

A. Image emotion inference.

Previous research has been devoted to inferring emotions from different types of multimedia data, such as texts, images, voices and videos. The research efforts have primarily focused on investigating emotions, extracting effective features and utilizing diverse types of learning methods.

Regarding emotion inference, the first question is how to properly model emotions. Despite extensive psychological research and debates, there is still no consensus on how to model emotions [33]. According to existing theories, emotions can be modeled in various ways. In general, two of the most popular theories are the categorical theory and the dimensional theory. The categorical theory indicates that emotions can be classified into certain categories. One example is to classify emotions into positive ones and negative ones [8], [34]. Moreover, Ekman [35] found a high agreement across members of different cultures on selecting emotional labels that fit facial expressions. The expressions that he found to be universal include happiness, surprise, anger, disgust, fear and sadness. Works of inferring emotions according to these categories include [36], [37], [12], [38], [24], [39]. However, some emotions that are frequently conveyed in facial expressions seldom appear in social images. For example, it is easy to detect *surprise* from one's face, but on image social networks, it is very difficult to determine what type of image is "*surprising*". Another popular theory is the eight basic emotion categories proposed by Mikels *et al.* [40]. These eight basic emotion categories are defined from the international affective picture system (IAPS) [41], a database of pictures designed to provide a standardized set of pictures for studying emotion and attention. They determined that the top 4 negative emotion categories are fear, sadness, disgust and anger, whereas the top 4 positive emotion categories are awe, amusement, excitement and content. Works on inferring image emotions according to these categories include [7], [10], [42], [43]. We can see that the negative emotions of Ekman's theory are the same as Mikels', but Ekman's positive emotion (happiness) is subdivided into four categories. Hence, Mikels' emotions can be regarded as a detailed variation of Ekman's emotions. The dimensional theory considers emotions to be variables with fuzzy boundaries. Emotions are represented as coordinates in a two- or three-dimensional space, such as valence-arousal, valence-arousal-dominance (also called evaluation-activity-potency, pleasure-arousal-dominance) [44] and Plutchik's wheel [45]. Specifically, Valdez and Mehrabian [46] investigated emotional reactions to color hue, brightness and saturation using the pleasure-arousal-dominance model. Works on inferring image emotions according to dimensional theory include [47], [48], [49]. Furthermore, there is a difference between perceived emotion and induced emotion, which has been reported by Juslin and Laukka [50]. Perceived emotion refers to what the viewer thinks the author wants to express, and induced emotion refers to the emotion actually felt by the viewer. The underlying mechanisms are different, and measuring induced emotion is more difficult than measuring perceived emotion.

In terms of inferring emotions from images, different types of visual features have been proven to be associated with the emotional contents of images. Machajdik *et al.* [7] extracted features representing the color, texture, composition and content and selected their proper combination for different data sets. Wang *et al.* [51] focused on mining the interpretable aesthetic visual features directly affecting human emotional perception from the perspective of art theories. These features include figure-ground relationship, color pattern, shape and composition. Datta *et al.* [52] inferred the aesthetic quality of pictures from their visual content. They extracted visual features that can discriminate between aesthetically pleasing and displeasing images, such as wavelet-based texture, size and aspect ratio, region composition, shape convexity, and so forth. Siersdorfer *et al.* [8] transformed image features into discrete elements or terms and described image contents in almost the same way as text documents. Borth *et al.* [11] established a mid-level representation by constructing a concept detector library called SentiBank. Zhao *et al.* explored principles-of-art features, including balance, emphasis, harmony, variety and movement, as mid-level features [42] and jointly combined them with low-level features such as GIST and high-level features such as facial expressions [5]. In the specific situation of social networks, in addition to the attribute correlation, which means the correlation between image emotions and visual features, more types of correlations are leveraged to help improve the accuracy. Jia *et al.* [9] not only utilized color features but also utilized the social correlation among images to infer emotions from social images, indicating that the emotion of current image may be associated with the emotion of the previous image that the same user uploads. Yang *et al.* [37] studied the social effect on image emotions and made full use of friend interactions, such as friends' comments on images.

For learning methods, traditional machine learning methods such as naive Bayesian were employed in [7]. Dellagiacoma *et al.* [36] chose the support vector machine (SVM) framework for the supervised learning of different emotion classes. In [9], [12], probabilistic graphical models were utilized to learn image emotions. In [5], a single graph was constructed for each type of feature, and multiple graphs were combined to learn the optimized weights of each graph to explore the complementation of different features in a regularization framework. For the joint modeling of images and texts, [34] used a deep neural network to analyze the visual-textual sentiment. Yang *et al.* [37] regarded visual features as a mixture of Gaussian, treated the corpus of comments as a mixture of topic models and integrated them using a cross-sampling process.

In addition to the three aforementioned aspects, there are other related works about image emotion inference. Wang *et al.* [12] verified the existence of emotion influence in image networks and discussed how social influence plays a role in changing users' emotions. Yang *et al.* [39] reported that the ability of emotionally influencing others is closely associated with users' social roles in image social networks, such as opinion leaders, structural hole spanners and ordinary users. Peng *et al.* [53] changed image emotion by using an emotion predictor. Retrieving images based on their emotional contents was studied in [2], [3], [4], [5].

Emotions can be inferred from other types of data besides images. In terms of texts, Calix *et al.* [54] recognized emotions in texts and used them to render facial expressions. In terms of voices, Tawari and Trivedi [14] explored the role of contextual information for speech emotion recognition. Luengo *et al.* [15] analyzed the characteristics of features derived from prosody, spectral envelope, and voice quality and their capability to discriminate emotions and validated them through experiments. In terms of music, Juslin and Laukka [50] provided an up-to-date overview of theory and research concerning expression, perception, and induction of emotion in music. For audio-visual signals, Mower *et al.* [17] studied the interaction between emotional audio and video cues. Lin *et al.* [19] presented an error-weighted semi-coupled hidden Markov model to recognize human emotions. Ozkan *et al.* [18] proposed a method based on a concatenated hidden Markov model (co-HMM) to infer both dimensional and continuous emotion labels from audio-visual cues. Wang *et al.* [20] introduced a kernel cross-modal factor analysis method into an audio-visual-based bimodal emotion recognition problem. Wu *et al.* [49] removed the speaking effect on facial expressions to improve the accuracy of emotion recognition. They also addressed the complex temporal course in face-to-face natural conversation in emotional expression [23]. Jiang *et al.* [55] proposed a comprehensive computational framework for predicting emotions carried by user-generated videos. Wang *et al.* [24] modeled the higher-order relations among emotions and proposed a multiple emotional multimedia tagging approach. They demonstrated the effectiveness of the proposed method on music, video and film databases.

In this paper, in contrast to other works that apply Ekman's [35] six emotion categories or Mikels' [40] eight emotion categories directly, we first consider how to model emotions of social images and reveal nine major emotional categories. In addition, we propose an effective model that considers visual features, correlations, and so forth in a joint framework to infer emotional tags from social images.

B. User demographics

In general, user demographics refer to the personal attributes of users. Different attributes are applied under different circumstances. In [56], user demographics contain gender, age, religion and political view on Facebook[3]. In [57], user demographics refer to users' age groups on mobile phone social networks. In [26], user demographics consist of gender, age and location. In [58], user demographics are composed of the neighborhood diversity and education levels.

Recently, research has verified the correlation between the demographics and behaviors of users on networks. Parag *et al.* [59] revealed that people who talk to each other are more likely than random to be of similar age and location. Bi *et al.* [56] discovered that the demographics of users are associated with their search query histories. Brea *et al.* [57] found a correlation between user demographics (more specifically, the users age) and the structure of their mobile phone social network. They summarize it as age homophily, indicating that contacts between similar people occur at a higher rate than between dissimilar people. Bakhshi *et al.* [58] leveraged user demographics to perform online restaurant recommendations, and Zhao *et al.* [60] made product recommendations based on matching the user demographic information extracted from their public profiles with product demographics learned from microblogs and online reviews.

Knowing user demographics is very helpful when personalizing web search results, query suggestions or recommen-

[3] http://www.facebook.com, a popular social network website.

dations. Although some behavioral and psychological studies have revealed correlations between human emotion perception and their demographics [28], [29], [30], [31], [32], whether a correlation exists between the emotional tags of social images and user demographics is still unclear. In this paper, we observe and validate the correlation between emotional tags of social images and user demographics. We further leverage the above findings into a partially labeled factor graph model to help improve the accuracy of inferring emotional tags from social images.

III. Problem Definition

In this paper, we study how user demographics are associated with emotional tags of social images and utilize them to improve the accuracy of inferring emotional tags. We provide definitions and formulate the problem in this section.

User demographics: In this paper, we present user v_i's demographics as three-dimensional vectors $\mathbf{p_i}$: *gender*, *marital status* and *occupation*.

Gender is defined as *male* or *female*. Marital status is defined as *single* or *married*. For occupation, because there are very many types of occupations, it is difficult to observe the difference of emotional tags among all occupations. To be consistent with gender (male and female) and marital status (single and married), we select *artist* and *engineer* as representatives. Ten occupations related with artist and 15 occupations related with engineer are selected. In detail, artist, writer, musician, dancer, photographer, film maker, designer, blogger, editor and freelancer are regarded as artist. Electrical / mechanical / biochemical / structural / civil / hydraulic / software engineers, programmer, web designer, network administrator, electrician, machinist, technician, architect, and scientist are regarded as engineer.

Emotional tag: The emotional tag of image $x^t_{i,j}$ uploaded by user v_i at time t is denoted as $y^t_{i,j}$, where j is the index of images uploaded by user v_i. We infer one emotional tag for one image. The emotional space is denoted as R.

In this work, the emotional tag is defined as perceived emotion, which means the emotion that the viewer thinks the owner wants to express, rather than the induced emotion, which means the emotion actually felt by the viewer.

Image social network: A partially labeled time-varying image social network can be defined as $G = (V, P, E^t, X^L, X^U)$. V is the set of users. P is the set of user demographics. E^t is the following relationship among users at time t. For instance, $e = (v^t_i, v^t_j) \subset E^t$ means that at time t, user v_i follows user v_j. X^L represents images whose emotional tags are available for training, and X^U represents images whose emotional tags are unavailable for training and only available for testing.

Based on the above definitions, the learning task of our work is formulated as follows.

Learning task: Given a partially labeled time-varying image social network $G = (V, P, E^t, X^L, X^U)$, find a function f to predict emotional tags from images:

$$f : G = (V, P, E^t, X^L, X^U) \rightarrow Y \tag{1}$$

where $Y = \{y^t_{i,j}\} \in R$.

TABLE I
Examples of Emotional Adjectives of Different Emotion Categories.

Emotion	Num.	Examples
awe	69	respectful, awesome, noble, worshipful, ...
amusement	35	amusing, funny, laughable, jokey, ...
contentment	100	delicious, delighted, cozy, cheerful, ...
excitement	69	vigorous, incredible, fascinating, superb, ...
anger	216	enraged, annoyed, furious, irritating, ...
disgust	147	distasteful, humiliating, loathsome, nasty, ...
sadness	155	depressed, disastrous, tearful, upset, ...
fear	174	afraid, dread, horrified, scary, ...
boredom	146	absentminded, alone, jejune, unattractive, ...

[1] *Num.* Number of adjectives.

TABLE II
The Image Number of Every Emotion Category.

anger	disgust	sadness	fear	boredom
10,251	11,363	12,777	12,879	13,388
awe	**amusement**	**contentment**	**excitement**	**total**
11,221	14,949	26,204	38,882	151,914

IV. The emotional social image data set

A. Data collection

We randomly download 2,060,353 images and 1,255,478 users' profiles from Flickr.com. The data set is denoted as $D = \{I, U\}$, where I refers to the image set and U refers to the user set.

When downloading images, only images uploaded by their owners are downloaded. Images shared by others from their owners are not considered. Given a user, we collect all images uploaded by him/her. The image contents contain views, objects, human, animals, and so forth. Each image is associated with its owner, title, description, tags, comments, url, taken time and shooting information (if present).

When downloading users' profiles, for each user, we collect his/her alias, location, list of users he/she has contact with, list of groups he/she participates in, homepage url, demographics information and self-introduction (if present).

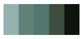
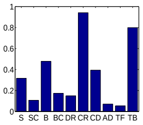
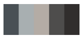
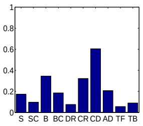
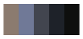
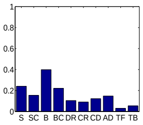
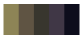
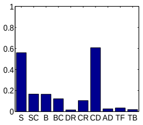
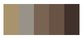
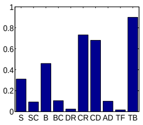
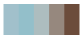
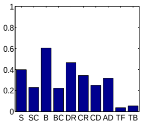
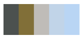
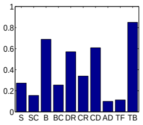
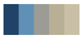
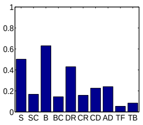
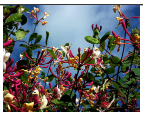
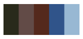
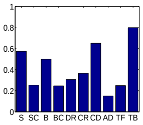
Because images and users are collected separately, for some users in U, images that they uploaded are not downloaded in I.

With the constraint that images must be associated with necessary information (owner, tags, and taken time), 854,734 images are qualified, which are uploaded by 3,181 users. On average, a user uploads 269 images. We denote this data set as $D_m = \{I_m, U_m\}$, where I_m refers to the image set and U_m refers to the user set. Because we obtain D_m by restricting the necessary information of D, D_m is the subset of D. In this data set, every user in U_m uploads images in I_m, and all images that they uploaded are also included in I_m.

B. Emotion categories

As discussed in related works, various types of theories can be applied to model emotions. Two of the most popular categorical theories are Ekman's six emotional categories [35] and Mikels' eight emotion categories [40].

TABLE III
EXAMPLES OF IMAGES AND THEIR VISUAL FEATURES FOR NINE EMOTION CATEGORIES.

Image	Visual feature		Image	Visual feature		Image	Visual feature	
Anger			Disgust			Sadness		
Fear			Boredom			Awe		
Amusement			Contentment			 Excitement		

In the specific situation of social networks, images uploaded and shared through the network may differ from images in real life. Based on Mikels' theory, we aim to find proper emotion categories for social images.

We select all the textual information of images, including title, description, tags and comments. We turn all the texts into lowercase and discard words that are less than two letters. Then, we use WordNet [61] to detect adjectives from the texts. For every adjective, we label it with one of the eight emotional categories. In detail, WordNet is a large lexical database of English. Words are grouped into sets of cognitive synonyms, each expressing a distinct concept. Given a pair of adjectives, it provides functions to calculate the semantic similarity between them. The greater the semantic similarity is, the greater the value that it returns. We choose *angry*, *disgusted/disgusting*, *sad*, *bored/boring*, *awed/awing*, *amused/amusing*, *content* and *excited/exciting* as core adjectives for the eight emotional categories. Then, given every adjective detected from the above texts, we calculate its emotional similarity between each of these core adjectives and label the adjective with the emotion category with the greatest semantic similarity. In this way, every adjective is labeled with the type of emotion that has the highest semantic similarity with it.

We observe the labeling results and discover the following:

- A substantial number of adjectives are semantically related with Mikel's eight emotion categories.
- However, a considerable number of adjectives such as "everyday", "our", "American", and "eleventh", which do not convey emotions semantically, are assigned to one of the eight categories due to the labeling process. Therefore, we discard these adjectives and consider them as *not emotional*.
- Moreover, there is a small group of adjectives such as "absentminded" and "unattractive" that are more closely associated with another type of emotion - *boredom*. After the image labeling process (which will be discussed in next subsection), we find that 8.81% of images are labeled *boredom*. Thus, we regard boredom as a new type of emotion.

With the above manual verification, we finally obtain nine emotion categories with relative emotional adjectives from social images on Flickr. For emotion categories, eight of them are defined by Mikels *et al.* [40]: *amusement*, *excitement*, *awe*, *contentment*, *disgust*, *anger*, *fear*, and *sadness*, and the ninth one is *boredom*. For relative emotional adjectives, examples and the number of adjectives for every emotion category are shown in Table I.

C. Automatic labeling

Due to the massive scale of our data set, manually labeling the emotional tag for every image is not practical. Herein, we adopt a strategy that is similar to the strategies used in [62] and [63] to label the emotional tags of images automatically, which is regarded as the ground truth.

In the above subsection, we define the emotion categories of social images. During the process, we obtain nine lists of emotional adjectives for nine emotion categories. Next, we expand the emotional adjective lists by adding the noun forms and verb forms of the adjectives (if present). Then, we compare the same texts selected above, including title, description and tags, with every word list, and an image is labeled with a type of emotional tag whose word list matches the words of tags most frequently.

In this way, 151,914 of the 854,734 images are labeled with emotional tags, which are uploaded by 2,300 users. On average, a user uploads 66 images with emotional tags. This data set is used in the following observations and experiments, which is denoted as $D_s = \{I_s, U_s\}$, where I_s refers to the image set and U_s refers to the user set. Because we obtain D_s by labeling image emotional tags from D_m, D_s is the subset of D_m. The number of images of every emotion category is summarized in Table II. Examples of images and their visual features are shown in Table III. For example, images of *amusement* tend to have high brightness and middle saturation [51].

V. OBSERVATIONS

The demographics of users have been verified to be associated with the behaviors of users in social networks [59],

TABLE IV
SUMMARY OF VISUAL FEATURES.

Abbr.	Dim.	Explanation
FC	15	Five dominant colors in HSV color space
S, SC	2	Saturation and its contrast
B, BC	2	Brightness and its contrast
DR	1	Dull color ratio
CR	1	Cool color ratio
CD	1	Color difference between foreground and background
AD	1	Area ratio difference between foreground and background
TF	1	Texture complexity of foreground
TB	1	Texture complexity of background

[1] *Abbr.* Abbreviation; *Dim.* Dimension

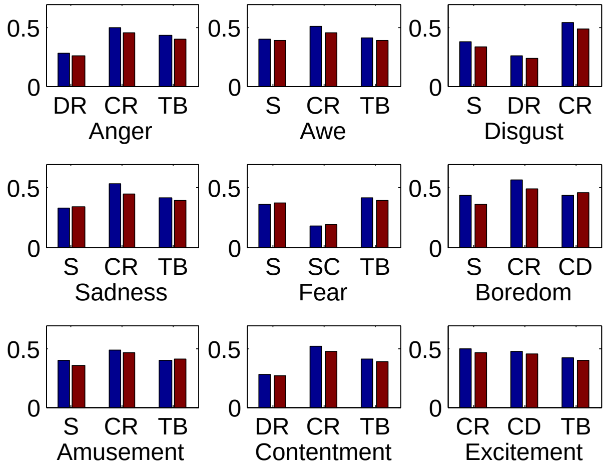


Fig. 2. The visual feature distributions of images uploaded by women (blue) and men (red).

[56], [57], [58], [26], [27]. Wondering whether a correlation between user demographics and emotional tags of social images exists on image social networks, we conduct a series of observations and uncover several phenomena.

Herein, we observe the correlation between emotional tags of images and three attributes of the user demographics. The data set that we employed is D_s, which consists of 151,914 images uploaded by 2,300 users and related metadata. A user uploads 66 images on average.

A. Visual features

We use functions in OpenCV[4] to convert images from RGB color space to HSV color space. HSV stands for hue, saturation, and value, and it is also often called HSB (B for brightness).

Then, we extract visual features according to the methods presented in [51].

S is the mean saturation of the image. In HSV space, a pixel has three channels: H, S and V. Channel S represents the saturation of the pixel. We scan every pixel of the image and sum the values of Channel S. Then, we divide the sum by the number of pixels and obtain the mean saturation of the image. *SC* is the mean saturation contrast of the image. Similarly, we scan every pixel of the image and sum the absolute values of Channel S. Then, we divide the sum by the number of pixels and obtain the mean saturation contrast of the image. *S* and *SC* describe the brilliant degree of colors and the differences in an image (*e.g.*, high saturation makes people feel fresh) [51].

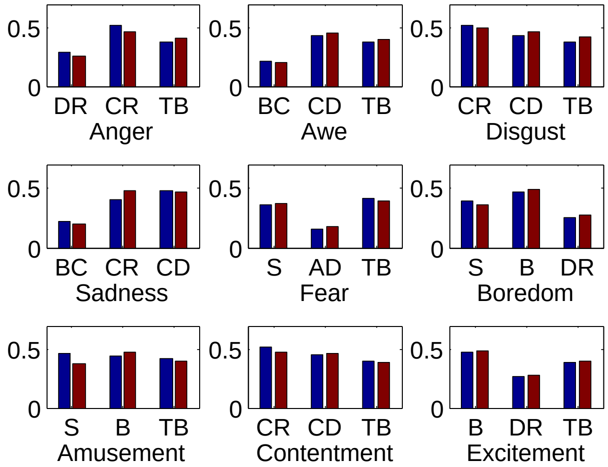


Fig. 3. The visual feature distributions of images uploaded by single users (blue) and married users (red).

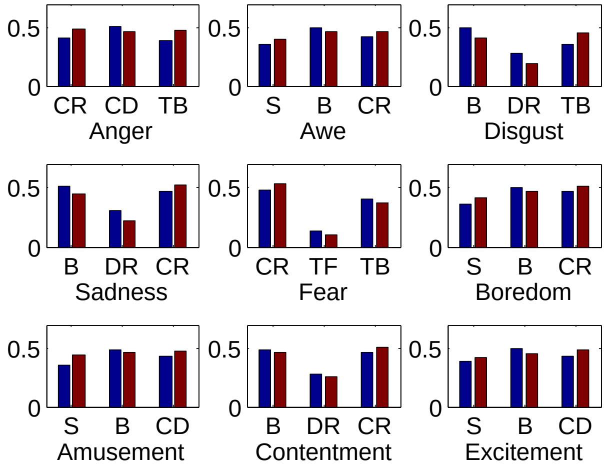


Fig. 4. The visual feature distributions of images uploaded by engineers (blue) and artists (red).

B is the mean brightness of the image, and *BC* is the mean brightness contrast of the image. In HSV space, Channel V represents the brightness of the pixel. We calculate *B* and *BC* using the same method for calculating *S* and *SC*. *B* and *BC* illustrate the black-while degree and the differences (*e.g.*, low brightness makes people feel negative and deep) [51].

DR stands for dull color ratio. A pixel is defined as “dull color” if its Channel V, which represents brightness, is greater than 0.7. We scan every pixel and calculate the ratio of “dull color” pixels.

CR stands for cool color ratio. A pixel is defined as “cool color” if its Channel H, which represents hue, is in the range of 30 and 110. We scan every pixel and calculate the ratio of “cool color” pixels. Cool colors such as blue and green make people calm, and warm colors such as red and yellow

[4] Open Source Computer Vision, a library of programming functions mainly aimed at real-time computer vision, http://opencv.org

can arouse excitement.

Then, for every image, we adopt the salient region detection technique to extract the foreground [64], [65]. Using this approach, the image is divided into two parts: foreground and background. We calculate the mean HSV values of the foreground and background. Then, *CD*, which stands for color difference, is calculated from the Euclidean distance of HSV values of foreground and the background. We also calculate the pixel numbers of the foreground and background. Then, *AD*, which stands for area difference, is calculated from the ratio difference of the foreground and background. These features describe the contrast between the foreground and background. *TF* and *TB*, which stand for the texture complexity of the foreground and background, are the density of Canny edges in the foreground and background, respectively.

The dimensions and explanation of these visual features are summarized in Table IV.

In total, 25 dimensions of visual features are extracted. The values are normalized between 0 and 1 over the entire data set.

The effectiveness of these features in inferring emotional tags from images has been confirmed in [9], [37], [12], [38]. Compared with low-level features such as SIFT and wavelet textures, these features are mid-level interpretable aesthetic features, which are more understandable for ordinary users.

B. Observation on the gender correlation.

We classify users into two groups according to their gender. If a user does not fill in his / her gender, we discard this user. This left us with 305 male users and 1,403 female users.

Then, we select images uploaded by these two groups of users and analyze the visual feature distributions of these images. Because there are many visual features, we only report the three most significant visual features that make the greatest difference between two groups in Fig. 2. The X-axis represents visual features, and the Y-axis represents the values of visual features, which have been normalized between 0 and 1.

We can observe that in the case of *boredom*, the visual feature distributions of images uploaded by men and women are different. For instance, the saturation (*S*) of images uploaded by women (0.4331) is 18.8% higher than that of images uploaded by men (0.3646). The cool color ratio (*CR*) of images uploaded by women (0.5614) is 15.6% higher than that of images uploaded by men (0.4855). This result suggests that although both men and women want to express their boredom through images, they use different visual features to convey their feelings.

In terms of *anger*, the cool color ratio (*CR*) of images uploaded by women (0.5025) is 11.6% higher than that of images uploaded by men (0.4504), and the dull color ratio (DR) of images uploaded by women (0.2753) is 8.5% higher than that of images uploaded by men (0.2548), which indicates that men and women have different ways of expressing their anger.

From the observation results, it can be concluded that there is a gender difference in the emotion tagging of social images.

C. Observations on the marital status correlation.

Similarly, according to the user's marital status, we divide users into single and married, each containing 259 and 825 users, respectively. If a user does not fill in his / her marital status, we discard this user. We conduct observations again, and the results are presented in Fig. 3.

The visual feature distributions of images uploaded by single users and married users are different. For example, in terms of *anger*, the cool color ratio (*CR*) of images uploaded by single users (0.5219) is 13.6% higher than that of images uploaded by married users (0.4595), and the texture complexity of the background (*TB*) of images uploaded by single users (0.3782) is 7.9% lower than that of images uploaded by married users (0.4081).

In terms of *sadness*, the color color ratio (*CR*) of images uploaded by single users (0.4007) is 19.7% lower than that of images uploaded by married users (0.4798), and the brightness contrast (*BC*) of images uploaded by single users (0.2158) is 7.4% higher than that of images uploaded by married users (0.2009).

The results show that single users and married users use different ways to assign emotional tags to social images.

D. Observations on the occupation correlation.

Because there are many types of occupations, it is difficult to observe the difference of emotional tags among all occupations. As described in Section III, we carefully select 25 types of occupations and classify them into "engineer" and "artist". "Artists" include writers, musicians, dancers, photographers, designers, and so forth, and "engineers" include programmers, mechanics, scientists, and so on. Then, we obtain 191 users as engineers and 234 users as artists. If a user does not fill in his / her occupation or the occupation is not included these 25 types, we discard this user. We conduct observations again, and Fig. 4 presents the results.

In terms of *disgust*, the brightness (*B*) of images uploaded by engineers (0.4957) is 19.4% higher than that of images uploaded by artists (0.4150). The dull color ratio (*DR*) of images uploaded by engineers (0.2790) is 44.6% higher than that of images uploaded by artists (0.1930). The texture complexity of the background (*TB*) of images uploaded by engineers (0.3560) is 28.7% lower than that of images uploaded by artists (0.4580).

In terms of *contentment*, the cool color ratio (*CCR*) of images uploaded by engineers (0.4654) is 9.3% lower than that of images uploaded by artists (0.5089). The brightness (*B*) of images uploaded by engineers (0.4886) is 4.6% higher than that of images uploaded by artists (0.4672).

The results suggest that on image social networks, engineers and artists have different patterns of emotional tags.

The observations can be summarized as follows:

- Men and women have different ways to assign emotional tags to social images, particularly *boredom* and *anger*. There is a gender difference in the emotion tagging of social images.
- Single users and married users use different ways to assign emotional tags, particularly *anger* and *sadness*,

indicating a marital status difference in the emotion tagging of social images.
- Engineers and artists use different patterns to assign emotional tags for most of the nine types of emotions, such as *disgust* and *contentment*, suggesting that occupation may play a key role in users' emotion tagging.

VI. MODEL

In this paper, we propose a demographics factor graph model (**D-FGM**) to leverage the above findings to help infer emotional tags from social images.

A factor graph is one type of probabilistic graphical model, and it provides an elegant way to represent both undirected graphical structures and directed graphical structures, with more emphasis on the factorization of the distribution [66] [1]. The basic idea of the factor graph model is to define correlations using different types of factor functions, and the objective function is defined based on the joint probability of the factor functions; thus, the problem of inferring emotional tags is cast as learning model parameters that maximize the joint probability.

In D-FGM, four types of correlations are defined as factor functions.

- **Visual feature correlation** $f_1(\mathbf{u}_{i,j}^t, y_{i,j}^t)$. It represents the correlation between visual features $\mathbf{u}_{i,j}^t$ and the emotional tag $y_{i,j}^t$.
- **Temporal correlation** $f_2(y_i^{t'}, y_i^t)$. Previous research has verified that there is a strong dependency between one's current emotion and the emotions in the recent past on social networks [1], [12]. This correlation is defined as temporal correlation, which represents the influence of the user's previous emotional tags in the recent past t' on the current emotional tag at time t.
- **Social correlation**. Creating and sharing images on image social networks is very different from traditional creation. Some users may have a strong influence on their friends' emotions, and some emotions may spread quickly on the social network [9], [37], [12]. In this paper, the social correlation contains three parts: the correlation between the emotional tag and the number of the user's friends $f_3(s_i^t, y_{i,j}^t)$, the correlation between the emotional tag and the major emotional tag of the user's friends $f_4(m_i^t, y_{i,j}^t)$ and the correlation between the emotional tag and the user's intimacy with friends $f_5(y_i^t, y_j^t, \mu_{i,j}^t)$.
- **User demographics correlation** $f_6(\mathbf{p}_i, y_{i,j}^t)$. It denotes the correlation between the emotional tag and the user's demographics information $\mathbf{p}_i$, which has been discussed in Section V. $\mathbf{p}_i$ is a three-dimensional vector: *gender*, *marital status* and *occupation*.

A. The predictive model

The notations of the proposed model are summarized in Table V. As described in Section III, the input of the model is an image social network G, and the output of the model is the inference results Y. The correlations described above are instantiated as different factor functions.

TABLE V
NOTATIONS IN THE PROPOSED MODEL.

Symbol	Description
$\mathbf{u}_{i,j}^t$	the set of visual features
$\mathbf{p}_i$	the demographics information of user v_i
s_i^t	the number of user v_i's friends
m_i^t	the major emotional tag of user v_i's friends
$\mu_{i,j}^t$	the intimacy between user v_i and user v_j at time t
$y_{i,j}^t$	the emotional tag of image $x_{i,j}^t$
λ	learning ratio
Z	normalization term
S	the aggregation of factor functions over all nodes
θ	parameter set

(1) **Visual feature correlation function**:

$$f_1(\mathbf{u}_{i,j}^t, y_{i,j}^t) = \frac{1}{z_\alpha} \exp\{\alpha^T \cdot \mathbf{u}_{i,j}^t\} \tag{2}$$

where $\mathbf{u}_{i,j}^t$ is the set of visual features and α is the parameter vector, indicating the weight of different visual features. All images share the same α.

(2) **Temporal correlation function**:

$$f_2(y_i^{t'}, y_i^t) = \frac{1}{z_\varepsilon} \exp\{\varepsilon_i \cdot g(y_i^{t'}, y_i^t)\}, t' < t \tag{3}$$

where y_i^t and $y_i^{t'}$ represent the emotional tag of user v_i at times t and t', respectively. Function $g(y_i^{t'}, y_i^t)$ is used to depict the correlation. ε_i is a one-dimensional parameter, indicating the weight of the temporal correlation of v_i. Images of v_i share the same ε_i.

(3) **Social correlation function**:

$$f_3(s_i^t, y_{i,j}^t) = \frac{1}{z_\gamma} \exp\{\gamma^T \cdot s_i^t\} \tag{4}$$

where s_i is the number of user v_i's friends and γ is a one-dimensional parameter, indicating the weight of s_i.

$$f_4(m_i^t, y_{i,j}^t) = \frac{1}{z_\delta} \exp\{\delta^T \cdot m_i^t\} \tag{5}$$

where m_i^t is the major emotional tag of user v_i's friends and δ is a one-dimensional parameter, indicating the weight of m_i^t. To calculate m_i^t, for an image that is uploaded by user v_i at time t, we observe the emotional tags of v_i's friends in the recent past (*e.g.,* 1 day). If v_i's friend v_j uploads an image in the recent past and v_i leaves a comment, v_i witnesses v_j's image and may be influenced by user v_j. Thus, we calculate the frequency of the emotional tags of user v_i's friends in the recent past, and we find the most frequent emotional tag as the major emotional tag of v_i's friends.

$$f_5(y_i^t, y_j^t, \mu_{i,j}^t) = \frac{1}{z_\eta} \exp\{\eta_{i,j} \cdot h(y_i^t, y_j^t, \mu_{i,j}^t)\} \tag{6}$$

where $\mu_{i,j}^t$ is the intimacy between user v_i and user v_j at time t. Function $h(y_i^t, y_j^t, \mu_{i,j}^t)$ is used to depict the correlation. $\eta_{i,j}$ is a one-dimensional parameter, indicating the weight of the social correlation between v_i and v_j. Images of v_i and v_j share the same $\eta_{i,j}$. To calculate $\mu_{i,j}^t$, for a pair of users v_i and v_j, we calculate their interaction

Algorithm 1 The learning and inference algorithm of emotional tags from social images.

```
Input:
  A partially labeled time-varying image social network $G = (V, P, E^t, X^L, X^U)$ and the learning ratio $\lambda$
Output:
  Construct a partially labeled factor graph.
  Initiate parameters $\theta = \{\alpha, \beta, \gamma, \delta, \varepsilon_i, \eta_{i,j}\}$
  repeat
    Calculate $E_{(P_\theta(Y|Y^U, G))}S$ using a standard algorithm Loopy Belief Propagation
    Calculate $E_{(P_\theta(Y|G))}S$ using a standard algorithm Loopy Belief Propagation
    Calculate the gradient of $\theta$: $E_{(P_\theta(Y|Y^U, G))}S - E_{(P_\theta(Y|G))}S$
    Update $\theta$ with learning ratio $\lambda$: $\theta = \theta_0 + \frac{\partial \mathcal{O}}{\partial \theta}\lambda$
  until convergence
  Obtain the inference results $Y = y_{i,j}^t$, $y_{i,j}^t \in R$ and the trained parameters $\theta = \{\alpha, \beta, \gamma, \delta, \varepsilon_i, \eta_{i,j}\}$
```

TABLE VI
THE F1-MEASURE OF 4 METHODS FOR 9 EMOTION CATEGORIES.

Emotion	NB	SVM	FGM	D-FGM
amusement	0.2028	0.0723	0.3569	**0.3851**
anger	0.1541	0.1897	0.2848	**0.2922**
awe	0.0042	0.0996	0.2777	**0.2916**
boredom	0.1187	0.1437	0.2423	**0.2657**
contentment	0.0984	0.0956	0.2188	**0.2458**
disgust	0.2432	0.2663	0.4055	**0.4246**
excitement	0.1544	0.1470	0.2418	**0.2588**
fear	0.0359	0.0987	0.2385	**0.2770**
sadness	0.1900	0.1849	0.2836	**0.3120**
average	0.1335	0.1442	0.2833	**0.3059**

TABLE VII
THE ABBREVIATIONS OF DIFFERENT FEATURE COMBINATIONS.

All	all features
-G	not include gender (a dimension of factor $f6$)
-M	not include marital status (a dimension of factor $f6$)
-O	not include occupation (a dimension of factor $f6$)
-D	not include user demographics (factor $f6$)
-fE	not include the major emotional tag of user's friends (factor $f4$)
-fS	not include the number of user's friends (factor $f3$)
-fI	not include the emotional impact of user's friends (factor $f5$)
-t	not include temporal information (factor $f2$)

frequency. Herein, the interaction means leaving a comment under a friend's image. We regard the frequency as three levels: *rarely*, *sometimes* and *often*. If the interaction frequency is under 10, it is regarded as *rarely*. If the interaction frequency is over 20, it is regarded as *often*. Otherwise, it is regarded as *sometimes*.

(4) **User demographics correlation function**:

$$f_6(\mathbf{p}_i, y_{i,j}^t) = \frac{1}{z_\beta} \exp\{\beta^T \cdot \mathbf{p}_i\} \tag{7}$$

where $\mathbf{p}_i$ is the demographics information of user v_i and β is a parameter vector, indicating the weight of different types of demographics (gender, marital status and occupation).

All parameters in the above functions are randomly initialized. Given the above factor functions, we define the joint distribution of the model. The joint distribution is the multiplication of factor functions over all images.

$$\begin{aligned} P(Y|G) = \frac{1}{Z} \prod_{x_{i,j}^t} f_1(\mathbf{u}_{i,j}^t, y_{i,j}^t) \prod_{x_{i,j}^t} \prod_{y_i^{t'}} f_2(y_i^{t'}, y_i^t) \\ \prod_{x_{i,j}^t} f_3(s_i^t, y_{i,j}^t) \prod_{x_{i,j}^t} f_4(m_i^t, y_{i,j}^t) \prod_{x_{i,j}^t} \prod_{v_j} f_5(y_i^t, y_j^t, \mu_{i,j}^t) \\ \prod_{x_{i,j}^t} f_6(\mathbf{p}_i, y_{i,j}^t) = \frac{1}{Z} \exp\{\theta^T S\} \end{aligned} \tag{8}$$

where $Z = Z_\alpha Z_\varepsilon Z_\beta Z_\gamma Z_\delta Z_\eta$ is the normalization term, S is the aggregation of factor functions over all nodes, and θ denotes all the parameters, i.e., $\theta = \{\alpha, \beta, \gamma, \delta, \varepsilon_i, \eta_{i,j}\}$.

Therefore, the target of modeling is to maximize the log-likelihood objective function $\mathcal{O} = \log P(Y|G)$.

B. Model learning

The objective function can be rewritten as:

$$\begin{aligned} \mathcal{O} = \log P(Y|G) = \log \sum_{Y|Y^U} \exp\{\theta^T S\} - \log Z \\ = \log \sum_{Y|Y^U} \exp\{\theta^T S\} - \log \sum_{Y} \exp\{\theta^T S\} \end{aligned} \tag{9}$$

Thus, the gradient of θ can be represented as:

$$\begin{aligned} \frac{\partial \mathcal{O}}{\partial \theta} = \frac{\partial(\log \sum_{Y|Y^U} \exp\{\theta^T S\} - \log \sum_{Y} \exp\{\theta^T S\})}{\partial \theta} \\ = E_{P_\theta(Y|Y^U, G)}S - E_{P_\theta(Y|G)}S \end{aligned} \tag{10}$$

In the above function, the first term is the expectation of inference results Y given social network G and Y^U, which is the inference results of X^U. The second term is the expectation of inference results Y given social network G, Y^U and Y^L, which are the inference results of X^U and X^L. X^L represents images whose emotional tags are available for training, and X^U represents images whose emotional tags are unavailable for training and only available for testing. Given the input and the output, we detail the learning process and summarize the algorithm in Algorithm 1.

The algorithm updates the parameters by $\theta = \theta_0 + \frac{\partial \mathcal{O}}{\partial \theta} \cdot \lambda$. The learning ratio λ is manually tuned. $\theta = 0.1$.

VII. EXPERIMENTS

A. Experimental setup

1) Data set: The data set is introduced Section IV. Herein, we use D_s for the experiments, which contains 151,914 emotional images uploaded by 2,300 users. To examine the

TABLE VIII
THE PRECISION, RECALL, F1-MEASURE AND ACCURACY OF BINARY CLASSIFICATION OF EVERY EMOTION CATEGORY.

	anger	**disgust**	**sadness**	**fear**	**boredom**	**awe**	**amusement**	**contentment**	**excitement**
Precision	0.6974	0.7378	0.6766	0.7362	0.7085	0.7356	0.7007	0.6840	0.6834
Recall	0.6284	0.5510	0.7223	0.5555	0.6072	0.4438	0.4412	0.5860	0.4714
F1-Measure	0.6611	0.6308	0.6987	0.6332	0.6540	0.5536	0.5415	0.6312	0.5580
Accuracy	0.6857	0.6794	0.6895	0.6809	0.6849	0.6415	0.6253	0.6597	0.6282

TABLE IX
THE PRECISION, RECALL AND F1-MEASURE OF D-FGM WITH 8 FEATURE COMBINATIONS FOR 9 EMOTION CATEGORIES. IN EACH COLUMN, RED AND BLUE COLORS REPRESENT THE BEST AND SECOND BEST FEATURE COMBINATIONS.

		amusement	**anger**	**awe**	**boredom**	**contentment**	**disgust**	**excitement**	**fear**	**sadness**
Precision	All	0.4261	0.3691	0.6338	0.2338	0.1652	0.5256	0.4386	0.3188	0.2986
	-G	0.3362	0.3314	0.5894	0.2155	0.1633	0.4367	0.3884	0.2899	0.2773
	-M	0.3072	0.3246	0.4773	0.2164	0.1671	0.5633	0.3372	0.2763	0.2512
	-O	0.3285	0.3314	0.4676	0.2097	0.1671	0.4126	0.3382	0.3072	0.2628
	-D	0.7217	0.3556	0.4937	0.2155	0.1671	0.2976	0.3671	0.2802	0.2599
	-E	0.3285	0.3324	0.4657	0.2126	0.1652	0.4145	0.3469	0.3005	0.2647
	-S	0.3246	0.3285	0.4696	0.2087	0.1671	0.4145	0.3362	0.2976	0.2618
	-I	0.3285	0.3275	0.4638	0.2097	0.1671	0.4126	0.3382	0.3024	0.2628
Recall	All	0.3507	0.2415	0.1894	0.3082	0.4889	0.3556	0.1836	0.2454	0.3266
	-G	0.3787	0.2473	0.1865	0.3043	0.3633	0.3826	0.1768	0.2184	0.3024
	-M	0.4213	0.2512	0.1961	0.2870	0.3314	0.3324	0.1826	0.2097	0.3130
	-O	0.3894	0.2531	0.1990	0.2899	0.3150	0.3990	0.1894	0.1990	0.3111
	-D	0.2928	0.2377	0.2029	0.2986	0.3758	0.4145	0.1855	0.2271	0.3179
	-E	0.3903	0.2502	0.1990	0.2879	0.3227	0.3971	0.1874	0.2019	0.3072
	-S	0.3884	0.2512	0.1942	0.2889	0.3227	0.3981	0.1913	0.1778	0.3121
	-I	0.3913	0.2522	0.1981	0.2870	0.3179	0.3971	0.1884	0.1971	0.3082
F1-Measure	All	0.3851	0.2922	0.2916	0.2657	0.2458	0.4246	0.2588	0.2770	0.3120
	-G	0.3556	0.2831	0.2831	0.2522	0.2251	0.4077	0.2425	0.2493	0.2889
	-M	0.3556	0.2831	0.2783	0.2473	0.2222	0.4184	0.2367	0.2386	0.2792
	-O	0.3565	0.2870	0.2792	0.2435	0.2184	0.4058	0.2425	0.2415	0.2850
	-D	0.4164	0.2850	0.2879	0.2502	0.2319	0.3469	0.2464	0.2512	0.2860
	-E	0.3565	0.2860	0.2783	0.2444	0.2184	0.4058	0.2435	0.2415	0.2841
	-S	0.3536	0.2850	0.2744	0.2415	0.2193	0.4058	0.2435	0.2232	0.2841
	-I	0.3565	0.2850	0.2783	0.2425	0.2193	0.4048	0.2415	0.2386	0.2841

performance of every emotion category, we evenly and randomly select 10,000 images from every emotion category. As for multiple classification, 90,000 images are chosen in total: 60% for training and 40% for testing. As for binary classification, for every emotion category, its 10,000 images are chosen as positive cases, and the other 10,000 images that are randomly selected from other eight emotion categories are chosen as negative cases.

2) Comparison methods: To demonstrate the effectiveness of our method, three learning methods, namely, naive Bayesian (NB), support vector machine (SVM) and traditional factor graph model (FGM), are chosen as baseline methods. We conduct comparison experiments on the same data set.

NB: Naive Bayesian is a widely used classifier and achieves good performance [7]. It is also used as the baseline method in [1]. We use the naive Bayesian tool provided by MATLAB[5].

SVM: SVM is a frequently used method in many classification problems. The method is also used as the baseline method in [36], [1], [9]. Herein, we use LIBSVM design by Chang and Lin [67].

FGM: This method is used in [9] to infer the emotions of images. A partially labeled factor graph model is utilized as a classifier.

D-FGM: D-FGM refers to the proposed method demographics factor graph model.

3) Evaluation metrics: We compare the performance of our proposed model with three baseline methods in terms of accuracy[6], precision[7], recall[6] and F1-measure[8]. These evaluation metrics are widely used in retrieval problems.

Note that as for multiple classification, the model will classify the image into one of the nine emotion categories. Thus, for an image whose true emotional tag is happiness, only when the inference result is happiness will it be calculated as true positive.

B. Experimental results

1) Multiple classification: Fig. 5 shows the mean accuracy. Table VI summarizes the F1-measure. Table VII summarizes the abbreviations of different feature combinations.

We can see that our model significantly enhances the performance. The accuracy achieves 0.2984, showing a +0.126 improvement compared with naive Bayesian (0.1724), a +0.1382 improvement compared with SVM (0.1602) and a +0.0167 improvement compared with FGM (0.2817). The average F1-measure reaches 0.3059, showing a +17.2% improvement compared with naive Bayesian, a +16.2% improvement compared with SVM and a +2.3% improvement compared with FGM.

[5] A widely used software developed by MathWorks, Inc.

[6] https://en.wikipedia.org/wiki/Accuracy_and_precision

[7] https://en.wikipedia.org/wiki/Precision_and_recall

[8] https://en.wikipedia.org/wiki/F1_score

In [37], a similar Flickr data set is used for multiple classification. They classify images into six emotion categories defined by Ekman [35]. Thus the problem defined in this manuscript is more difficult to deal with because it has more emotion categories as candidates. However, the proposed method still shows a satisfying performance. For example, the F1-Measure of *disgust* reaches 0.4246, which is +0.0506 higher than [37] (0.3740.)

2) Binary classification: The precision, recall, F1-Measure and accuracy are shown in Table VIII. The proposed method shows a competitive result by showing the average accuracy of 0.6639.

In this manuscript, eight of the emotions we find on Flickr are defined by Mikels *et al.* [40] and the ninth one is *boredom*. Though the nine emotion categories are first used in this manuscript, the eight emotion categories defined by Mikels *et al.* [40] are widely used in previous research, including [7], [11], [42], [43], *etc*. Though different algorithms are proposed and different data sets are employed in these works, the results reported by above research can be used for reference. The true positive rate shown in [7] on a combined data set is 0.60. (They use true positive rate per class averaged over the positive and negative classes instead of the correct rate over all samples.)

3) Analysis: NB and SVM are only capable of handling *vectors*. For social images, the visual features, user demographics and parts of the social attributes (the number of the user's friends and the major emotional tag of the user's friends) are modeled as vectors in the input. However, these two methods cannot handle the correlations between images, which are instantiated as *edges* in FGM and D-FGM. Consequently, they neglect the temporal correlation and the intimacy with the user's friends, which negatively impacts the performance.

For FGM, it can model the vectors and edges jointly. However, all the edges are modeled with the same weight in FGM. Thus, although the method can model the temporal correlation, it cannot model the users' intimacy with friends, in which case the intimacy is modeled as the weight of the edge in the input. This constraint negatively impacts the performance.

In contrast, the proposed D-FGM can model the vectors, edges and weighted edges in a joint framework; thus, it considers all the information of social images and achieves the best performance.

Simply, we find that the F1-measure of the proposed method is relatively high when inferring *disgust*. Regarding *disgust*, it has been reported to be a prototypic emotion that encompasses a variety of reaction patterns according to the subjective experiences of different individuals [68], [37]. Thus, taking the information of user demographics into consideration is very important.

C. Factor contribution analysis

In our work, we utilize the information of the user demographics and introduce them into a factor graph model as factor functions. Wondering whether these factors benefit the inference, we investigate the contribution of every factor in the model.

The precision, recall and F1-measure of D-FGM with 8 feature combinations for 9 emotion categories are shown in Table IX. Table VII summarizes the abbreviations of different feature combinations. Each time, we take each of the factors out of the primitive model and examine the performance while the other factors remain the same. For example, in "-G", we take gender out of the features while the others remain the same. In "-M", we take marital status out of the features while the others remain the same.

From Table IX, we can see that irrespective of whether NB, SVM, FGM or the proposed D-FGM is applied, the feature combination involving all factors achieves the best performance in most emotion categories. The results validate the effectiveness of the factors.

Specifically, to examine the contribution of user demographics in D-FGM, we visualize the results with all factors / all factors but user demographics in polar coordinates. As shown in Fig. 6, the origin of polar coordinates represents 0 in terms of F1-measure. Every dot represents a type of emotional tag, and the length between the dot and the origin of polar coordinates shows the F1-measure of this emotional tag.

The interesting results are summarized as follows.

- When inferring *anger*, the gender information benefits the inference (+3.2% improvement).
- When inferring *fear*, *sadness*, *excitement* and *awe*, the marital status information has a considerable impact. The F1-measure increases by 16.3%, 11.9%, 9.6$ and 5.0%, respectively.
- When inferring *contentment*, *boredom* and *disgust*, the occupation information is very useful by showing improvements of 12.6%, 9.3% and 4.5%, respectively.
- Interestingly, however, when inferring *amusement*, the demographics information does not help, which indicates that the pattern of tagging *amusement* is primarily determined by the visual features.

The results also correspond to the observation results, which verifies the rationality of introducing the demographics of the users into the modeling of inferring emotional tags from social images.

D. Case study

In the above investigation, we observe that different user demographics result in different patterns of emotional tags of images. Herein, we detail the analysis by reporting the emotional tags, the visual features and the user demographics of several images in Table X.

The two images on the left depict a similar scene, and their visual features are quite similar. However, we find that the image on the top is uploaded by a female on April, 14th, 2011, who tags this image as *excitement*, and the image on the bottom is uploaded by a male on June, 16th, 2010, who tags this image as *boredom*. The gender difference in human emotion perception is verified by the behavioral study [28].

Similarly, the two images in the middle both capture sharp rocks and streams, but the top one expresses *boredom* by a single user on December, 14th, 2010, and the bottom one conveys *contentment* by a married user on May, 8th, 2011,

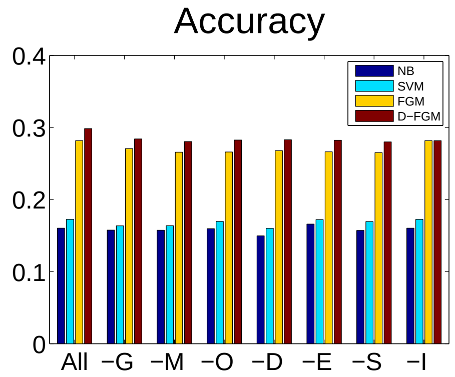


Fig. 5. The mean accuracy of 4 methods.

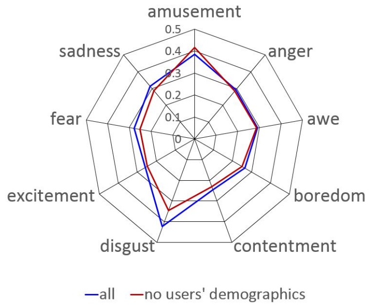


Fig. 6. The F1-measure of D-FGM with 1) all factors and 2) all factors but user demographics.

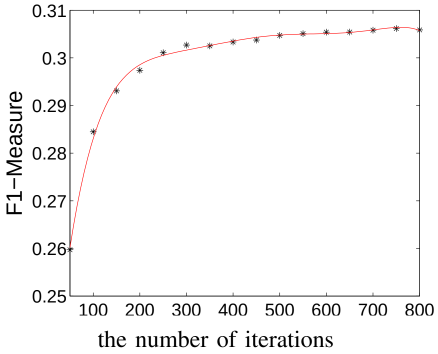
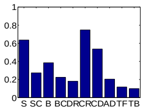
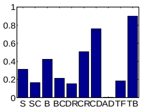
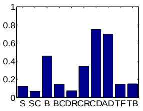
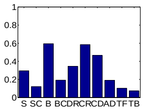
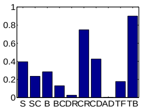
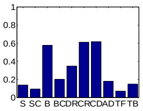


Fig. 7. F1-measure of the proposed D-FGM.

TABLE X
DIFFERENT USER DEMOGRAPHICS RESULTS IN DIFFERENT EMOTIONAL TAGS OF IMAGES.

Image & Emotional tag written by the owner	**Visual features**	**User demo-graph-ics**	**Image & Emotional tag written by the owner**	**Visual features**	**User demo-graph-ics**	**Image & Emotional tag written by the owner**	**Visual features**	**User demo-graph-ics**
 Excitement		**Female**	 Boredom		Male **Single**	 Awe		Female **Artist**
 Boredom		**Male**	 Contentment		Male **Mar-ried**	 Sadness		Male **Engi-neer**

indicating that single users and married users have different patterns of emotional tags of social images.

The two images on the right both depict a shining sea. However, the image on the top is taken to express *awe* by an artist on July, 9th, 2010, and the image on the bottom is taken to express *sadness* by an engineer on October, 27th, 2010. The results demonstrate the different patterns of emotional tags between engineers and artists.

E. Parameter Analysis

In this section, we report the evolution of the F1-measure as the number of iteration increases. In every iteration, the proposed D-FGM makes inferences for every image and optimizes its parameters. When the number of iterations is 50, the average F1-measure of the nine emotion categories is 0.2597. As the number of iterations increases, the F1-measure increases as well. When the number of iterations is greater than 700, the F1-measure reaches convergence. The average F1-measure of the nine emotion categories is 0.3059 when the number of iterations is 800.

F. Error Analysis

Finally, we would like to present our analysis on the possible sources of errors based on the inference results of the proposed D-FGM.

1) Noise and missing data: To evaluate the performance of the proposed D-FGM, we first have to know the primitive emotional tags of social images. However, the amount of social images is incredibly large. Thus, manually labeling the emotional tag of every image is not practical. In this paper, we adopt an automatic labeling process to determine the primitive emotional tags of social images. In this process, we consider all the textual information written by the image owners, including titles, tags and descriptions. This idea is simple and practical, but it also introduces some noise. First, not all images are associated with the above textual information. Second, the words written by image owners may be not sufficiently accurate.

2) Other factors: Inferring emotional tags is a very difficult task because emotions are highly subjective and complicated. Occasionally, different types of feelings can mix together, and these can be called new types of emotional tags. At present, there is still no consensus on how to model emotions. In this paper, we adopt the basic eight emotion categories proposed by Mikels [40] and add a newly observed one - “boredom”. These categories may not cover all the human feelings that users want to express on image social networks.

VIII. CONCLUSIONS

In this paper, we study the problem of “link inferring with user demographics” for inferring emotional tags from social images. First, we consider how to measure the emotions of social images and exploit nine major emotion categories. Then, we investigate whether user demographics are related to the emotional tags of social images and unveil several interesting patterns. By introducing these patterns as factor functions into modeling, we propose a demographics factor graph model (**D-FGM**), which infers emotional tags from social images not only by visual features, temporal correlation and social correlation but also by user demographics. Experiments on the world's largest image sharing website Flickr validate the effectiveness of the proposed method.

For our future work, more information can be taken into consideration. For example, users from different regions and culture backgrounds may use different patterns to assign emotional tags. In addition, because users' emotions may be influenced by friends, when considering the emotion influence on social networks, we can not only utilize the pairwise emotion influence but also explore the emotion influence of different social structures (triangle, rectangle, and so forth).

Regarding applications, one the one hand, understanding users' emotions can help images be retrieved not only based on their contents but also their emotions. On the other hand, understanding users' emotions can help virtual personal assistants (*Siri*[9], *Cortana*[10], *Google Now*[11] and so forth) to humanize their responses and build better human-mobile/human-computer interactions. As virtual personal assistants are becoming increasingly more popular, users expect to communicate with virtual personal assistants not only by instructions and queries but also through chats and conversations. It is important that virtual personal assistants be able to understand users' intentions and emotions. If the user is very angry and he/she sends an image with an angry face, the virtual personal assistant may first apologize and then generate new responses.

Acknowledgment

This work is supported by National Key Research and Development Plan (2016YFB1001200), the Innovation Method Fund of China (2016IM010200). This work is supported by the National Natural Science Foundation of China (61370023, 61521002, 61561130160, 61602033, 61561130160). This work is partially supported by the major project of the National Social Science Foundation of China(13&ZD189).

References

[1] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. Fong, "Quantitative study of individual emotional states in social networks," *IEEE Transactions on Affective Computing (TAC)*, vol. 3, pp. 132–144, 2012.

[2] W. Wei-Ning, Y. Ying-Lin, and J. Sheng-Ming, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *IEEE International Conference on Systems, Man and Cybernetics*, 2006, pp. 3534–3539.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *Acm Computing Surveys*, vol. 40, no. 2, p. 2007, 2008.

[4] K. A. Olkiewicz and U. Markowska-Kaczmar, "Emotion-based image retrieval - an artificial neural network approach." in *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)*, 2010, pp. 89–96.

[5] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1025–1028.

[6] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.

[7] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92.

[8] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *the 18th ACM international conference on Multimedia*, 2010, pp. 715–718.

[9] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang, "Can we understand van gogh's mood? learning to infer affects from images in social networks," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 857–860.

[10] B. Li, S. Feng, W. Xiong, and W. Hu, "Scaring or pleasing: exploit emotional impact of an image," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1365–1366.

[11] D. Borth, R. Ji, T. Chen, T. Breuel, and S. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 223–232.

[12] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, and L. Xie, "Modeling emotion influence in image social networks," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 286–297, 2015.

[13] L. Zhang, M. Song, N. Li, J. Bu, and C. Chen, "Feature selection for fast speech emotion recognition," in *International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, October*, 2009, pp. 753–756.

[14] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, 2010.

[15] I. Luengo, E. Navas, and I. Hernaez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, 2010.

[16] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[17] E. Mower, M. J. Mataric, and S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Transactions on Multimedia*, vol. 11, 2009.

[18] D. Ozkan, S. Scherer, and L. P. Morency, "Step-wise emotion recognition using concatenated-hmm," in *ACM International Conference on Multimodal Interaction*, 2012, pp. 477–484.

[19] J. C. Lin, C. H. Wu, and W. L. Wei, "Error weighted semi-coupled hidden markov model for audio-visual emotion recognition." *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 142–156, 2012.

[20] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[21] Y. Song, L. P. Morency, and R. Davis, "Learning a sparse codebook of facial and body microexpressions for emotion recognition," in *ACM on International Conference on Multimodal Interaction*, 2013, pp. 237–244.

[22] J. C. Levesque, L. P. Morency, and C. Gagne, "Sequential emotion recognition using latent-dynamic conditional neural fields," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–6.

[23] C. H. Wu, J. C. Lin, and W. L. Wei, "Two-level hierarchical alignment for semi-coupled hmm-based audiovisual emotion recognition with temporal course," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1880–1895, 2013.

[24] S. Wang, J. Wang, Z. Wang, and Q. Ji, "Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 1–1, 2015.

[25] L. Pang, S. Zhu, and C. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1–1, 2015.

[9] http://www.apple.com/ios/siri/

[10] http://www.microsoft.com/en-us/mobile/campaign-cortana/

[11] http://www.google.com/landing/now/

[26] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 15–24.

[27] H. Huang, J. Tang, S. Wu, L. Liu, and X. Fu, "Mining triadic closure patterns in social networks," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 499–504.

[28] A. Fischer, P. Mosquera, A. Vianen, and A. Manstead, "Gender and culture differences in emotion." *Emotion*, vol. 4, no. 1, pp. 87–94, 2004.

[29] R. W. Simon and L. E. Nath, "Gender and emotion in the united states: Do men and women differ in selfreports of feelings and expressive behavior?" *American Journal of Sociology*, vol. 109, no. 5, pp. 1137–1176, 2004.

[30] J. W. Pennebaker and T. A. Roberts, "Toward a his and hers theory of emotion: Gender differences in visceral perception," *Journal of Social and Clinical Psychology*, vol. 11, no. 3, pp. 199–212, 1992.

[31] J. Crawford, S. Kippax, J. Onyx, U. Gault, and P. Benton, *Emotion and Gender: Constructing Meaning from Memory*. Sage Publications, 1992.

[32] R. Fivush, M. A. Brotman, J. P. Buckner, and S. H. Goodman, "Gender differences in parentchild emotion narratives," *Sex Roles*, vol. 42, no. 3, pp. 233–253, 2000.

[33] A. Ortony and T. Turner, "Whats basic about basic emotions?" *Psychological Review*, vol. 97, no. 3, pp. 315–331, 1990.

[34] Q. You, J. Luo, H. Jin, and J. Yang, "Joint visual-textual sentiment analysis with deep neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1071–1074.

[35] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[36] M. Dellagiacoma, P. Zontone, G. Boato, and L. Albertazzi, "Emotion based classification of natural images," in *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, 2011, pp. 17–22.

[37] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, "How do your friends on social media disclose your emotions?" in *the 28th AAAI Conference on Artificial Intelligence*, 2014, pp. 306–312.

[38] B. Wu, J. Jia, Y. Yang, P. Zhao, and J. Tang, "Understanding the emotions behind social images: Inferring with user demographics," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, 2015, pp. 1–6.

[39] Y. Yang, J. Jia, B. Wu, and J. Tang, "Social role-aware emotion contagion in image social networks." in *the 30th AAAI Conference on Artificial Intelligence*, 2016.

[40] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior Research Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[41] P. Lang, M. Bradley, and B. Cuthbert, "International affective picture system (iaps): Affective ratings of pictures and instruction manual. technical report a-8."

[42] S. Zhao, Y. Gao, X. Jiang, H. Yao, T. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 47–56.

[43] M. Chen, L. Zhang, and J. P. Allebach, "Learning deep features for image emotion classification," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015.

[44] C. Osgood, G. Suci, and P. Tannenbaum, "The measurement of meaning," *Attitudes*, vol. 7, no. 7, p. 342, 1957.

[45] R. Plutchik, "A general psychocvolutionary theory of emotion," *Emotion: Theory, Research, and Experience*, vol. 1, pp. 3–31, 1980.

[46] P. Valdez and A. Mehrabian, "Effects of color on emotions." *Journal of Experimental Psychology General*, vol. 123, no. 4, pp. 394–409, 1994.

[47] P. Lee, Y. Teng, and T. Hsiao, "Xcsf for prediction on emotion induced by image based on dimensional theory of emotion," in *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation*, 2012, pp. 375–382.

[48] S. Zhao, H. Yao, and X. Jiang, "Predicting continuous probability distribution of image emotions in valence-arousal space," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 879–882.

[49] C. H. Wu, W. L. Wei, J. C. Lin, and W. Y. Lee, "Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1732–1744, 2013.

[50] P. N. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of New Music Research*, vol. 33, no. 3, pp. 217–238, 2004.

[51] X. Wang, J. Jia, J. Yin, and L. Cai, "Interpretable aesthetic features for affective image classification," in *ICIP*, 2013, pp. 3230–3234.

[52] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision*, 2006, pp. 288–301.

[53] K. C. Peng, K. Karlsson, T. Chen, and D. Q. Zhang, "A framework of changing image emotion using emotion prediction," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014.

[54] R. A. Calix, S. A. Mallepudi, B. Chen, and G. M. Knapp, "Emotion recognition in text for 3-d facial expression rendering," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 544–551, 2010.

[55] Y. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *the 28th AAAI Conference on Artificial Intelligence*, 2014.

[56] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel, "Inferring the demographics of search users: social data meets search queries," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 131–140.

[57] J. Brea, J. Burroni, M. Minnoni, and C. Sarraute, "Harnessing mobile phone social network topology to infer users demographic attributes," in *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, 2014.

[58] S. Bakhshi, P. Kanuparthy, and E. Gilbert, "Demographics, weather and online reviews: a study of restaurant recommendations," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 443–454.

[59] Parag, Singla, Matthew, and Richardson, "Yes, there is a correlation: -from social networks to personal behavior on the web," in *In WWW 08: Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 655–664.

[60] X. W. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li, "We know what you want to buy: a demographic-based system for product recommendation on microblogs," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1935–1944.

[61] C. Fellbaum and G. Miller, "Wordnet : an electronic lexical database," *Cognition Brain and Behavior*, 1998.

[62] L. Xie, "Picture tags and world knowledge," in *ACM Multimedia*, 2013.

[63] S. J. Hwang, "Discriminative object categorization with external semantic knowledge," 2013, ph.D. Dissertation, The University of Texas at Austin.

[64] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 409–416, 2015.

[65] S. M. Hu, T. Chen, K. Xu, M. M. Cheng, and R. R. Martin, "Internet visual media processing: A survey with graphics and vision applications," *Visual Computer*, vol. 29, no. 5, pp. 393–405, 2013.

[66] Z. H. Tang, W. and J. Tang, "Learning to infer social ties in large networks." in *In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'11)*, 2011, pp. 381–397.

[67] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machines," *Acm Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 389–396, 2011.

[68] J. Moll, R. D. Oliveira-Souza, F. T. Moll, F. A. Igncio, I. E. Bramati, E. M. Caparelli-Dquer, and P. J. Eslinger, "The moral affiliations of disgust," *Cognitive and Behavioral Neurology*, vol. 18, no. 1, pp. 68–78, 2005.

Boya Wu Boya Wu is currently a master candidate in Department of Computer Science and Technology of Tsinghua University, Beijing, China. Her research interests include affective computing and social network analysis.

Jia Jia Jia Jia is currently an associate professor in Department of Computer Science and Technology of Tsinghua University, Beijing, China. Her research interests include affective computing and human-computer speech interaction.

Yang Yang Yang Yang is an assistant professor at College of Computer Science and Technology, Zhejiang University. His research focuses on mining deep knowledge from large-scale social and information networks.

Peijun Zhao Peijun Zhao is currently a master candidate in Department of Computer Science and Technology of Tsinghua University, Beijing, China. His research interests include affective computing and human-computer speech interaction.

Jie Tang Jie Tang is currently an associate professor in Department of Computer Science and Technology of Tsinghua University, Beijing, China. His research interests include social network theories, data mining methodologies, machine learning algorithms, and semantic web technologies.

Qi Tian Qi Tian is currently a Full Professor in the Department of Computer Science, the University of Texas at San Antonio (UTSA). His research interests include multimedia information retrieval, computer vision, pattern recognition and bioinformatics.