This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2617741, IEEE Transactions on Multimedia

Continuous Probability Distribution Prediction of Image Emotions via Multi-Task Shared Sparse Regression

Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Guiguang Ding

***Abstract*—Previous works on image emotion analysis mainly focused on predicting the dominant emotion category or the average dimension values of an image for affective image classification and regression. However, this is often insufficient in various real-world applications, as the emotions that are evoked in viewers by an image are highly subjective and different. In this paper, we propose to predict the continuous probability distribution of image emotions which are represented in dimensional valence-arousal space. We carried out large-scale statistical analysis on the constructed Image-Emotion-Social-Net dataset, on which we observed that the emotion distribution can be well modelled by a Gaussian mixture model. This model is estimated by an expectation-maximization algorithm with specified initializations. Then we extract commonly used emotion features at different levels for each image. Finally, we formalize the emotion distribution prediction task as a shared sparse regression (SSR) problem and extend it to multi-task settings, named multi-task shared sparse regression (MTSSR), to explore the latent information between different prediction tasks. SSR and MTSSR are optimized by iteratively reweighted least squares. Experiments are conducted on the Image-Emotion-Social-Net dataset with comparisons to three alternative baselines. The quantitative results demonstrate the superiority of the proposed method.**

***Index Terms*—Image emotion, probability distribution, valence-arousal, Gaussian mixture model, shared sparse regression, multi-task learning**

I. Introduction

IMAGES can convey rich semantics and evoke strong emotions in viewers. Understanding the perceived emotions in images has been widely studied recently due to its vital importance in various applications, ranging from entertainment to education and advertisement [1]–[3]. However, it is not a trivial problem at all, mainly due to the great challenges of affective gap [2], [4] and subjective evaluation [2], [5]–[8]. On one hand, affective gap can be defined as "the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal" [4]. On the other hand, subjective evaluation refers to the fact that "people may have different evoked emotions on the same image due to the difference of social and cultural backgrounds" [2], [5]–[8].

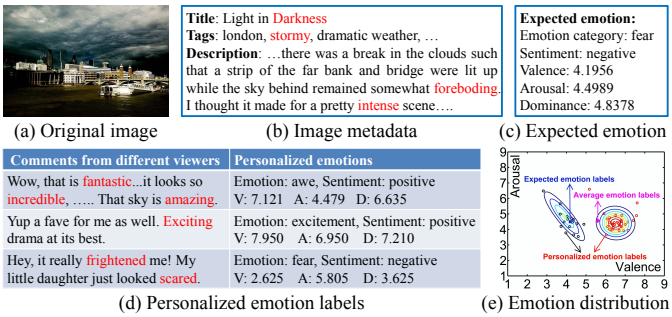


Fig. 1. The differences between traditional affective image regression and the proposed emotion distribution prediction. (a) is the uploaded image to Flickr. (b) are the title, tags and description given by the uploader to (a). (c) are the expected emotion labels that we assign to the uploader using the keywords in (b) in red. (d) are the comments to (a) from different viewers and the personalized emotion labels that we obtain using the keywords in red. (e) illustrates the differences, where the hollow points are the perceived emotion labels in VA space, the blue square and magenta diamond points are the target average VA scores by the traditional affective image regression methods using different strategies of obtaining labels, while the contour lines of GMM are the target emotion distribution by the proposed method.

Existing work mainly focused on finding features that can express emotions better to bridge the affective gap. To this end, such methods fall into the traditional classification or regression scenarios, trying to assign the dominant emotion category or the average dimension values to an image. However, predicting only the dominant emotion is insufficient in many applications, as the emotions that are evoked in different viewers by an identical image are highly subjective and different [7], due to various contextual factors, such as social and cultural influence [5]. For example, an image of stormy weather (Figure 1 (a)) may evoke feelings of excitement to some observers who like photographing of marvellous phenomenon, but fear in others who are afraid of thunder. Under such a circumstance, it is natural to take the subjective evaluation into account. This refers to predicting the personalized emotion perceptions for user-centric computing and predicting the probability distribution of image emotions for image-centric computing. To the best of our knowledge, there are few works

Manuscript received January 18, 2016; revised April 30, 2016 and August 9, 2016; accepted October 5, 2016. Date of publication October 15, 2016; date of current version October 15, 2016. This work was supported by the National Natural Science Foundation of China (No. 61571269, 61271394, 61472103, 61133003, 61671267, 61422210, 61373076). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marco Bertini.

S. Zhao, Y. Gao and G. Ding (corresponding author) are with the School of Software, Tsinghua University, Beijing 100084, China (e-mail: schzhao@gmail.com; gaoyue@tsinghua.edu.cn, dinggg@tsinghua.edu.cn).

H. Yao is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: h.yao@hit.edu.cn).

R. Ji is with the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, Xiamen 361005 China (e-mail: rrji@xmu.edu.cn).

tackling the subjective evaluation challenge on predicting the personalized perceptions as well as the probability distribution of image emotions.

In this paper, we make an initial attempt to predict the continuous probability distribution of image emotions represented in dimensional valence-arousal space, as shown in Figure 1. To accomplish this task, we construct a large-scale personalized image emotion dataset, named Image-Emotion-Social-Net, with images fully automatically crawled from Flickr. Related lexicon-based text emotions are viewed as personalized perceptions of image emotions. By the statistical analysis of personalized emotion perceptions on the Image-Emotion-Social-Net dataset, we observe that the valence-arousal emotion labels can be well represented by a Gaussian mixture model (GMM), i.e. a mixture of bidimensional Gaussian distributions. The expectation-maximization (EM) algorithm with specified initializations is then used to estimate the parameters of GMM. Subsequently, the emotion distribution prediction is formalized as a shared sparse regression (SSR) problem. SSR is then extended to multi-task setteings as multi-task shared sparse regression (MTSSR), which can predict the emotion distribution for multiple images simultaneously. Iteratively reweighted least squares is adopted to optimize SSR and MTSSR. Commonly used visual emotion features of three different levels are extracted. We conduct experiments on the constructed Image-Emotion-Social-Net dataset to demonstrate the effectiveness of the proposed method.

The contributions of this paper are three-fold:

1. Different from traditional methods for affective image regression, we propose to predict the continuous probability distribution of image emotions, which can be viewed as an initial attempt to tackle the subjective evaluation challenge.
2. (Multi-task) shared sparse regression together with three baseline methods are presented as learning models to predict the continuous probability distribution of image emotions. Iteratively reweighted least squares is used to optimize (MT)SSR.
3. We construct and release a large-scale emotion distribution dataset containing 18,700 images. The statistical analysis on this dataset further demonstrates the necessity of subjective evaluation. The experimental results on this dataset demonstrate that the proposed (MT)SSR outperforms several state-of-the-art methods.

One preliminary version on continuous probability distribution prediction of image emotions was first introduced in our previous work [9]. In this paper we have improvements in four aspects: (1) we perform a more comprehensive survey of related works; (2) we provide the detailed dataset construction process and more statistical analysis; (3) we improve the methods of SSR and MTSSR by adding constraints on the sparse coefficients to ensure that in predicted GMM the mixing coefficients sum to 1 and the covariance matrixes are positive definite; and (4) we conduct more comparative experiments and enrich the analysis of the results.

The remainder of this paper is organized as follows: Section II reviews related work. The motivation of this paper, including the constructed Image-Emotion-Social-Net dataset and its related statistical analysis, and the problem definition of emotion distribution prediction are described in Section III. We present the feature extraction and the emotion distribution prediction models, including the proposed (MT)SSR and three baseline methods in Section IV and V, respectively. Experimental evaluation and analysis are given in Section VI, followed by the conclusion and future work in Section VII.

II. Related Work

Image emotion and sentiment analysis. To analyze emotions from a given image, there are two widely used models: categorical emotion states (CES) and dimensional emotion space (DES). CES methods model emotions as one of a few basic categories [10]–[17], while DES methods employ 3-D or 2-D space to represent emotions, such as valence-arousal-dominance (VAD) [18], natural-temporal-energetic [19] and valence-arousal (VA) [2], [4], [13]. VAD is the most widely used DES, where valence represents the pleasantness of a stimulus ranging from happy to unhappy, arousal represents the intensity of emotion provoked by a stimulus ranging from excited to calm, while dominance represents the degree of control exerted by a stimulus ranging from controlled to in control [20], [21]. Specifically, image emotion is often called image sentiment for binary classification (positive or negative) [1], [3], [22], [23]. Accordingly, related work on image emotion analysis can be classified into three different tasks: affective image classification [1], [2], [11]–[17], [24], [25], regression [2], [13] and retrieval [10], [26]. Discrete probability distribution has been preliminarily investigated in [6], [7] based on CES models. We model image emotions using dimensional valence-arousal representations to predict continuous probability distributions.

From a feature's view point, visual features are designed and extracted of different levels, i.e., the different aspects or extends of image representations [1], [26]–[29]. Low-level holistic features including Wiccest features and Gabor features were extracted to classify image emotions in [12]. Machajdik *et al*. [11] extracted features inspired from psychology and art theory, such as *color*, *texture* and *composition*. Lu *et al*. [13] investigated the computability of emotion through *shape* features. Zhao *et al*. [2] proposed to extract more interpretable mid-level emotion features based on principles-of-art, such as *balance*, *contrast*, *harmony* and *variety*. Visual sentiment ontology and detectors are proposed to detect high-level adjective noun pairs based on large-scale social multimedia data [1], [3]. Yuan *et al*. [23] used mid-level scene attributes for binary sentiment classification. Simple social correlation features are explored for emotion classification of social network images [16]. As a special case of image emotion, facial expression recognition is also widely studied in recent years [30]–[36]. Tkalcic *et al.* [37] proposed to obtain the affective labels of images based on users' facial emotion expressions. We extract commonly used emotion features of three different levels [26] to test their performances for emotion distribution prediction.

Based on the extracted features, state-of-the-art methods tried to assign a dominant emotion category or the average

dimension values to an image for affective image classification and regression with CES and DES models, respectively. The commonly used models are based on machine learning methods, such as Naive Bayes [11], support vector machine (SVM) or support vector regression (SVR) [2], [13], sparse learning [7], [14], multi-graph learning [26] and convolutional neural network (regression) (CNN(R)) [6]. We present shared sparse regression for emotion distribution prediction and extend it to multi-task settings.

Note that affective content analysis has also been widely studied based on other types of input data, such as text [38], speech [39], [40], music [41]–[44] and videos [45]–[49].

Probability distribution prediction. In many applications of machine learning, it would be more reasonable and useful to predict the probability distribution for a target variable rather than simply the most likely value for that variable [50]. Probability distribution prediction has been studied in some areas, such as surf height [50], user behavior [51] and spike events [52]. According to probability theory, there are typically two types of probability distributions: discrete probability distribution and continuous probability distribution. Generally, the distribution prediction task can be formalized as a regression problem. For different emotion representation models, the distribution prediction varies slightly. For CES, the task aims to predict the discrete probability of different emotion categories, the sum of which is equal to 1. CNNR and shared sparse learning are recently used to predict the discrete probability distribution of image emotions in [6] and [7]. For DES, the task usually transfers to predict the parameters of specified continuous probability distribution, the form of which should be firstly decided, such as Gaussian distribution and exponential distribution. In this paper, we focus on the latter one, i.e., predicting the continuous probability distributions of image emotions.

Sparse learning and multi-task learning. Sparse learning represents the target variable as a sparsely linear combination of a set of basis functions and is widely used in many areas, such as face recognition [53], visual classification [54] and emotion analysis [7], [14].

Meanwhile, in many real-world applications, some classification/regression/clustering tasks may be related to each other [55]. For example, in the prediction of therapy outcome, the tasks of predicting the effectiveness of several combinations of drugs are related [56]. Traditional single-task learning methods solve these tasks independently, which ignores the task relatedness. Learning these tasks simultaneously by extracting and utilizing appropriate shared information across different tasks has been empirically [57], [58] as well as theoretically [59] proved to often significantly improve performances relative to single-task learning. A survey on multi-task learning can be referred to [55]. By combining sparse learning and multi-task learning, Wang *et al*. [60] proposed sparse multi-task regression for brain imaging identification. Similarly, we present multi-task shared sparse regression but with different optimization functions (using ℓ_0-norm instead of ℓ_1-norm) and constraints for emotion distribution prediction.

TABLE I
THE KEYWORD EXAMPLES OF EACH EMOTION CATEGORY. '#' INDICATES THE TOTAL KEYWORD NUMBERS.

Emotion	#	Keyword examples
amusement	24	amused, amusement, cheer, delight, funny, pleasing
anger	79	angry, annoyed, enraged, hateful, offended, provoked
awe	36	amazing, astonishment, awesome, impressive, wonderful
contentment	28	comfortable, gladness, happy, pleasure, satisfied
disgust	35	detestation, disgusted, nauseous, queasy, revolt, weary
excitement	49	adventure, enthusiastic, inspired, stimulation, thrilled
fear	71	afraid, frightened, nightmare, horror, scared, timorous
sadness	72	bereaved, heartbroken, pessimistic, sadness, unhappy

III. PROBLEM DESCRIPTION

In this section, we introduce the dataset (Image-Emotion-Social-Net[1]) on emotions of social images and describe the problem definition of continuous distribution prediction of image emotions.

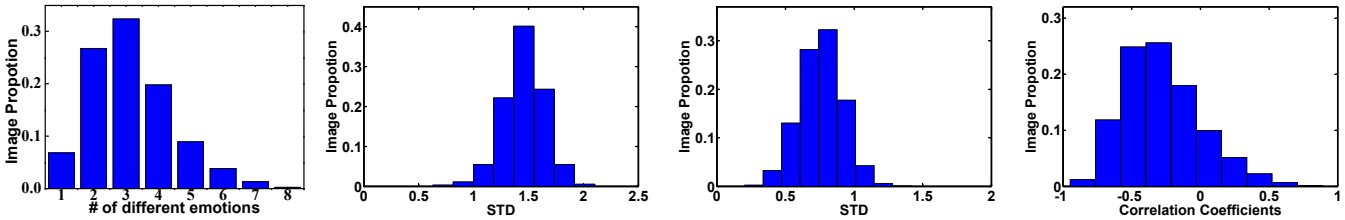
A. The Image-Emotion-Social-Net Dataset

1) Dataset Construction: We downloaded 21,066,920 images from Flickr with 2,060,357 users belonging to 264,683 groups. Each image is associated with the metadata, such as the title, tags, taken time and location if available. Each user is associated with the personal information, the contact list and the group list they joined in. As how to measure emotions is still far from consensus in research community [61], we defined emotions using both categorical and dimensional representations. For CES, we used the 8 categories rigorously defined in psychology [62], including 4 negative and 4 positive emotions. To get the ground truth labels, we adopted keywords based searching strategy as in [1], [16], [17]. Tens of keywords for each emotion category are obtained from a public synonym searching site[2] and are manually verified, with examples shown in Table I. Expected emotions of the image uploaders are firstly considered. The keywords are searched from the title, tags and descriptions given by the uploaders. The emotion category with the most frequent keywords is set as the ground truth of expected emotions from the uploaders, as shown in Figure 1 (b) and (c).

As we focus on personalized emotion perception, we then searched from all the comments of related images to get the actual emotion labels of each viewer. We removed the images if the searched title, tags or descriptions contain negation adjacent and prior to the target keywords, such as "I am not happy". Similarly, we also removed the comments with negation adjacent and prior to the target keywords. Note that the labels of an image for a specific user are allowed to have different emotion categories (such as fear, disgust) but must have only one sentiment (positive or negative). Then we computed the average value of valence, arousal and dominance as ground truth for dimensional emotion representation based on recently published VAD norms of 13,915 English lemmas [21], as shown in Figure 1 (d). Besides, we also gave the sentiment categories (positive or negative). We combined the expected emotions and actual emotions of all involved images for each

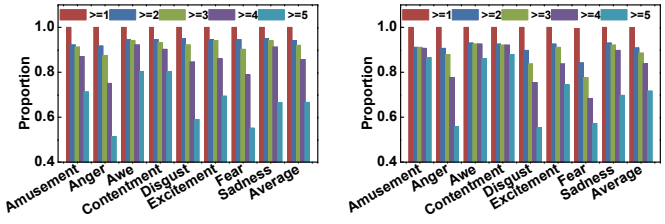
[1]https://sites.google.com/site/schzhao/
[2]http://www.thesaurus.com/browse/synonym/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2617741, IEEE Transactions on Multimedia



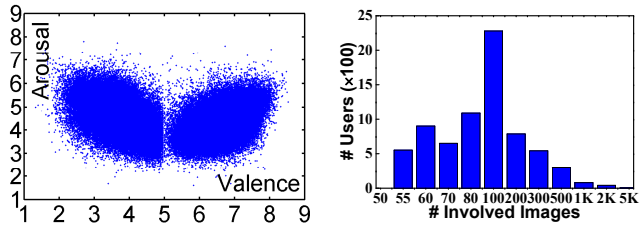
(a) Emotion number (b) STD of valence (c) STD of arousal (d) Correlation coefficients

Fig. 2. The statistics of CES and DES of the 18,700 images in the Image-Emotion-Social-Net dataset. We can find that only a tiny proportion of images convey just 1 emotion and the STD of valence for most images is larger than 1.5, which demonstrates the subjectiveness of emotion perceptions.



(a) Expected emotion (b) Personalized emotion

Fig. 3. Dataset validation results. $>= n$ means at least n Yes's. On average more than 88% of emotion labels receive at least 3 Yes's, which verifies the quality of the dataset.

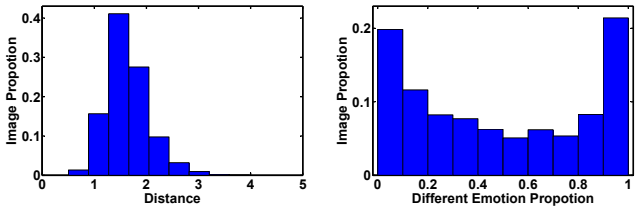


(a) VA distribution (b) User distribution

Fig. 4. Dataset statistics results on VAD distribution and user distribution. (a) is consistent with traditional emotion space [20]. (b) approximatively follows a typical Gaussian distribution.

TABLE II
IMAGE NUMBERS OF CATEGORICAL EMOTIONS.

amusement	awe	contentment	excitement	positive
270,748	328,303	181,431	115,065	1,016,186
anger	disgust	fear	sadness	negative
29,844	20,962	55,802	57,476	362,400



(a) VA distance (b) CES proportion

Fig. 5. The relation between expected vs personalized emotions of the images with more than 20 emotion labels on both DES and CES representations.

user. This process resulted in a dataset containing 1,012,901 images uploaded by 11,347 users and 1,060,636 comments on these images commented by 106,688 users. We chose 7723 active users with more than 50 involved images. Finally we obtained 1,434,080 emotion labels of three types, including 8 emotion categories, 2 sentiment categories and continuous values of valence, arousal and dominance. All the involved images of one user are labelled with sentiment categories and VAD values, while a tiny proportion of them are not assigned with the emotion categories if no keyword is found.

If one user is the uploader of an image, then the emotion of the metadata text (title, tags and descriptions) is the personalized emotion of this user, which is also the expected emotion that is expected to evoke in other viewers by this user. If one user is a viewer of an image, then the emotion of the comment is the personalized emotion of this user.

2) Dataset Validation: To validate the quality of the dataset, we did a crowdsourcing experiment on discrete emotions. For each emotion category, we randomly selected 200 images with associated titles, tags and descriptions for expected emotions, and 200 comments with corresponding images for personalized emotions. 5 graduate students (3 males, 2 females) were invited to judge whether the text was used to express the assigned emotions of related images. To facilitate this judgement, they were asked simple question like "*Do you think that the text is used to express excitement for this image?*", and they just needed to choose YES or NO. The result is shown in Figure 3. We can find that for both expected and personalized emotions, on average more than 88% of emotion labels receive at least 3 Yes's, which verifies the quality of the constructed dataset. In such cases, the expected emotion labels are 3.5% more accurately assigned than personalized emotions. To assess the inter-rater agreement, we also calculate the Fleiss' kappa[3] of the 5 annotators. The average Fleiss' kappa (the standard deviation) for the 8 emotion categories of expected emotions and personalized emotions are 0.2297 (0.0748) and 0.3224 (0.1411), respectively.

3) Statistics of Dataset: The distribution of images per emotion category is shown in Table II, where the first four columns represent the number of images in each of the 8 emotions; while the last column is the number of images with binary sentiments. We can find that the number of negative emotions is relatively small. The distribution of valence, arousal (without showing dominance here) is illustrated in Figure 4(a), which looks like a petal or heart, similar to the emotion space in [20]. The user distribution based on involved images is shown in Figure 4(b).

Totally we select 18,700 images with more than 20 VA labels each for the experiments on emotion distribution pre-

[3]https://en.wikipedia.org/wiki/Fleiss%27_kappa

diction. The distribution of emotion numbers for these images is shown in Figure 2(a). The histogram of valence and arousal standard deviations (STD) are shown in Figure 2(b) and Figure 2(c), while the histogram of the correlation coefficients is shown in Figure 2(d). Some image examples and related personalized emotion labels are shown in Figure 6(a) and 6(b). We can find that the emotion perceptions of different users are truly subjective and personalized.

We also analyze the relation between the expected and personalized emotions. For each of the images with more than 20 labels, we compute the Euclidean distances between personalized emotions and expected emotion in VA space, and average all the distances. The histogram of the average VA distance is shown in Figure 5(a). For CES, we count the proportion of personalized emotions that are different from expected emotion for each image. The histogram of different emotion proportions is illustrated in Figure 5(b). It is clear that there exists great inconsistency between expected and personalized emotions. Some image examples with high different emotion proportions are shown in Figure 7.

4) Challenging Tasks: The challenging tasks that can be performed by researchers on this dataset include, but not limited to, the following aspects:

- Image-centric emotion analysis. For each image, we can predict the dominant or expected emotion category like traditional affective image classification. Besides, we can predict the emotion distribution of each image, taking the normalized emotion proportion as the ground truth.
- User-centric emotion prediction. For each user, we can predict her personalized emotion perception of some specific images. The above two tasks can be extended to regression tasks, all of which can be done using visual, social, temporal and the combination of all features.
- Emotion related data mining and applications. This dataset contains visual and social information to support research on emotion influence mining, social advertising and affective image retrieval, *etc*.

For different tasks, the roles of expected emotions and actual emotions are different. For image-centric expected emotion analysis, only the expected emotions can be used. For image-centric dominant emotion analysis or emotion distribution analysis, the expected emotions can be viewed as one type of actual emotions. For user-centric emotion prediction, the expected emotions can also be viewed as one type of actual emotions, but only for the uploaders of related images.

In this paper, we focus on the image-centric emotion distribution analysis, trying to predict the continuous probability distribution of image emotions when perceived by large quantity of viewers, including the image uploaders.

B. Problem Definition

From Figures 6(a) and 6(b), we have the following observations: (1) The emotions evoked by an image in different viewers are truly subjective and different; Just assigning the average dimensional values of valence and arousal to an image is obviously not enough; (2) Though highly different, the perceived emotions follow certain distributions, which can be clearly grouped into two clusters, corresponding to the positive and negative sentiments; In each cluster, the VA emotion values are relatively stable; (3) The VA emotion labels can be well modeled by a mixture of two bidimensional Gaussian distributions.

Based on these observations, we define the distribution of VA emotion labels as a GMM by

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{l=1}^{L} \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{1}$$

where $\mathbf{x} = (v, a)$ is pair-wise VA emotion labels, $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ are the mean vector and covariance matrix of the lth Gaussian component, while π_l is the mixing coefficient, which satisfies $\pi_l \geq 0$ and $\sum_{l=1}^{L} \pi_l = 1$. In this paper, the number of Gaussian components is 2, i.e. $L = 2$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \pi_1, \pi_2)$. It should be noted that the number of Gaussian components L can be easily enlarged if more personalized emotion labels are obtained.

The EM algorithm is used to estimate the parameters of GMM. Specifically, the initializations are obtained by firstly partitioning the VA labels into two clusters based on whether valence is greater than 5, the typical boundary of positive and negative sentiments [21], and then computing the mean vector $\boldsymbol{\mu}_l$ and covariance matrix $\boldsymbol{\Sigma}_l$ of each cluster. The mixing coefficients are set as the proportions of related VA labels in each cluster to the total labels. In experiment, the EM algorithm is converged in 6.28 steps on average without overfitting. Some estimated results of GMM and detailed parameter values are shown in Figure 6(c).

Suppose we have N training images $\mathbf{f}_1, \cdots, \mathbf{f}_N$, the emotion distributions are $p_1(\mathbf{x};\boldsymbol{\theta}_1), \cdots, p_N(\mathbf{x};\boldsymbol{\theta}_N)$, where $\boldsymbol{\theta}_n = (\boldsymbol{\mu}_{n1}, \boldsymbol{\Sigma}_{n1}, \boldsymbol{\mu}_{n2}, \boldsymbol{\Sigma}_{n2}, \pi_{n1}, \pi_{n2})$ are the parameters of the nth emotion distribution $(n = 1, \cdots, N)$. Similarly, suppose we have M test images $\mathbf{g}_1, \cdots, \mathbf{g}_M$ with ground truth emotion distributions $q_1(\mathbf{x};\boldsymbol{\vartheta}_1), \cdots, q_M(\mathbf{x};\boldsymbol{\vartheta}_M)$, where $\boldsymbol{\vartheta}_m (m = 1, \cdots, M)$ are the distribution parameters. Then our goal is to predict the emotion distribution parameters $\widehat{\boldsymbol{\vartheta}}_m$ based on $\{\mathbf{f}_n, \boldsymbol{\theta}_n\}_{n=1}^{N}$ for each $\mathbf{g}_m$. That is

$$f : (\{\mathbf{f}_n, \boldsymbol{\theta}_n\}_{n=1}^{N}, \mathbf{g}_m) \rightarrow \widehat{\boldsymbol{\vartheta}}_m. \tag{2}$$

IV. Extracted Emotion Features

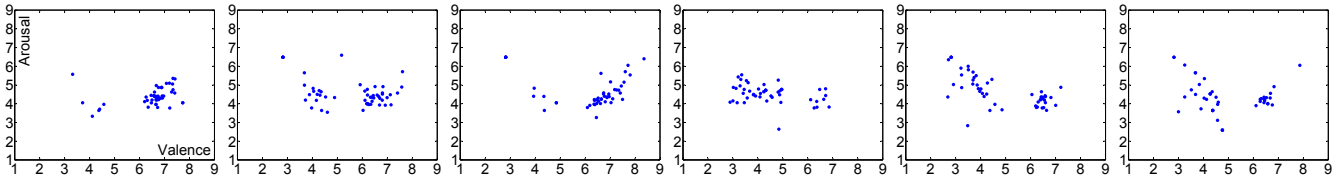
As shown in [26], there are various types of features that may contribute to the perceptions of image emotions. Similar to [26], we extract commonly used emotion features of different levels and generalities for each image, including low-level GIST [63] and elements-of-art [11], mid-level attributes [64] and principles-of-art [2], and high-level ANPs [1] and expressions [30].

Low-level features suffer from the difficulty of easy interpretation and the link to emotions is weak [26]. In this paper, we just extract GIST as **generic** feature, one of the most commonly used features, for its relatively powerful description ability of visual phenomena in a scene perspective [63], [64].

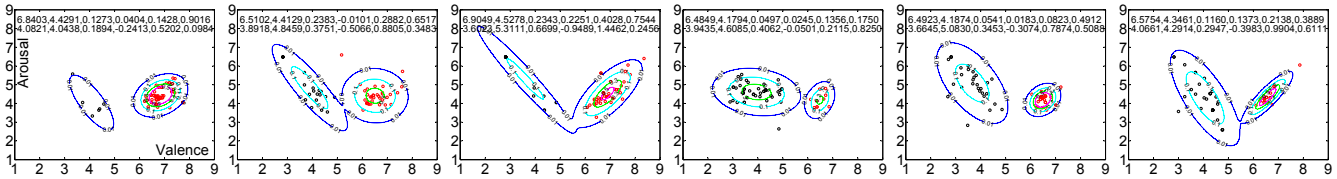
We extract **special** features derived from elements of art, including color and texture [11]. Low-level color features include mean saturation and brightness, vector based mean hue, emotional coordinates (pleasure, arousal and dominance)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2617741, IEEE Transactions on Multimedia

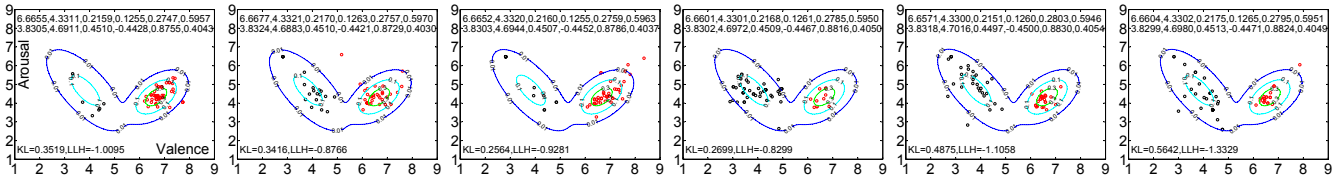
(a) Image examples in the Image-Emotion-Social-Net dataset



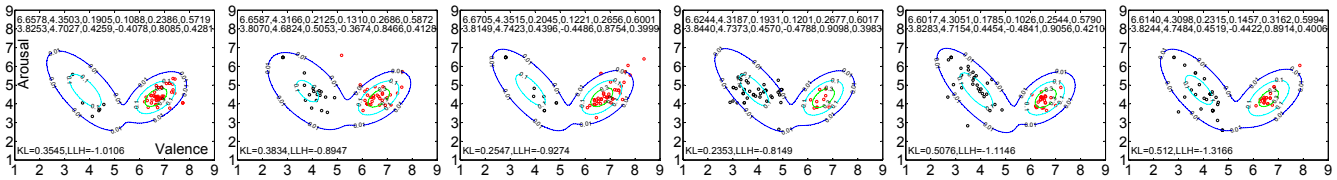
(b) Emotion distributions of personalized perceptions in VA space



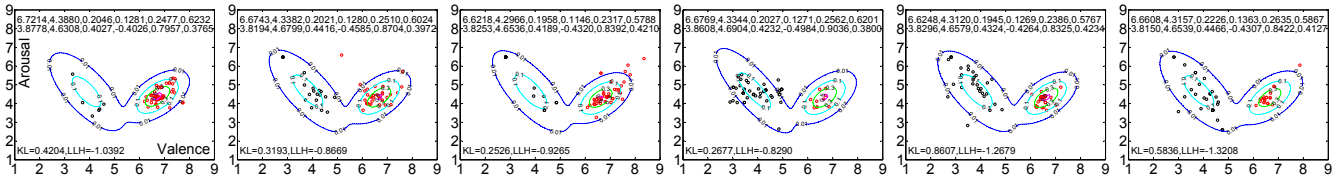
(c) The estimated GMM using specified EM algorithm



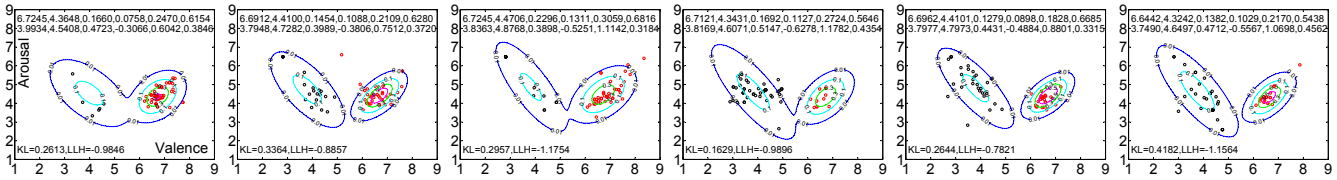
(d) The predicted GMM by GW using ANP features



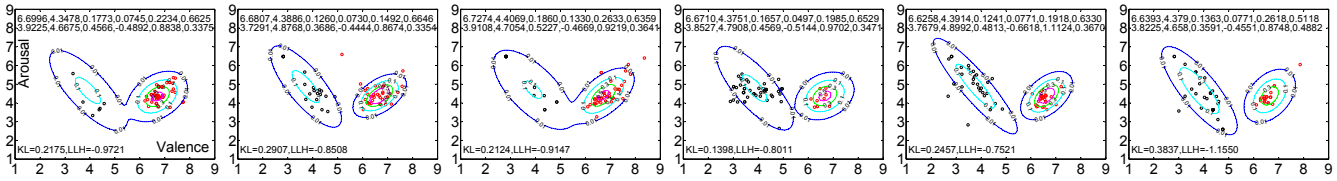
(e) The predicted GMM by KNNW ($K = 300$) using ANP features



(f) The predicted GMM by SVR using Elements features



(g) The predicted GMM by SSR using Principles features



(h) The predicted GMM by MTSSR using ANP features

Fig. 6. Image examples, related emotion labels and the predicted emotion distributions by different methods using related best features on the Image-Emotion-Social-Net dataset. The first and second line numbers in (c) to (h) are the GMM parameters corresponding to positive and negative sentiments, $\mu_{11}, \mu_{12}, \Sigma_{111}, \Sigma_{112}, \Sigma_{122}, \pi_1$ and $\mu_{21}, \mu_{22}, \Sigma_{211}, \Sigma_{212}, \Sigma_{222}, \pi_2$. It is clear that the proposed (MT)SSR outperforms the baselines on KL and LLH.

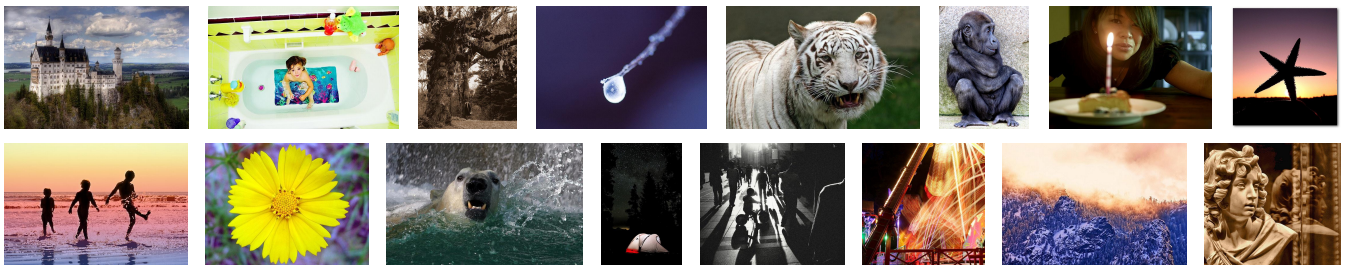


Fig. 7. Image examples with high different proportions (>0.9) between personalized and expected emotions, which indicates that the personalized emotion perceptions greatly differ from the expected emotions, probably due to different responses to the visual semantics (see the metadata and comments in Fig. 1).

based on brightness and saturation, colorfulness and color names. Low-level texture features include Tamura texture, Wavelet textures and gray-level co-occurrence matrix (GLCM) based texture [11].

Mid-level features are more semantic, more interpretable and have stronger link to emotions than low-level features [2]. Recently, attribute based representation has been widely studied for its intuitive interpretation and cross-category generalization property in visual recognition domain [64]–[66]. We extract 102 dimensional attributes which are commonly used by humans to describe scenes as mid-level **generic** features. As in [64], the attributes can be classified into five types: materials (mental), surface properties (dirty), functions or affordances (reading), spatial envelop attributes (cluttered) and object presence (flowers). GIST features and SVM implemented in Liblinear toolbox[4] are used to train attribute classifiers based on 14,340 images in SUN database [63].

Features inspired from principles of art, including balance, contrast, harmony, variety, gradation, and movement are extracted as mid-level **special** features [2]. These artistic principles are used to arrange and orchestrate artistic elements in art theory for describing specific semantics and emotions and are proved to have stronger link to emotions than elements. Please refer to [2] for detailed implementations.

High-level features are the detailed semantic contents contained in images. People can easily understand the emotions conveyed in images by recognizing the semantics. Concepts described by 1,200 adjective noun pairs (ANPs) [1] are extracted as **generic** features. The ANPs are detected by a large detector library SentiBank [1], which is trained on about 500k images downloaded from Flickr using various low-level features, including GIST, color histogram, LBP descriptor, attribute, *etc*. Liblinear SVM is used as classifier by early fusion. Finally, we obtain a 1,200 dimensional vector describing the probability that each ANP is detected.

Motivated by the conclusion that facial expressions may determine the emotions of the images containing faces [26], we also extract 8 kinds of facial expressions (*anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, *neutral*) [67] as high-level **special** features. Compositional features of local Haar appearance features are built by a minimum error based optimization strategy, which are embedded into an improved AdaBoost algorithm [30]. Trained on CK+ database [67], the method performs well even on low intensity expressions [30]. Face detection is firstly conducted using the Viola-Jones algorithm [68] to decide whether an image contains faces. Finally, we can get a 8 dimensional vector, each of which represents the proportion of related facial expressions.

[4]http://www.csie.ntu.edu.tw/~cjlin/liblinear/

TABLE III
SUMMARY OF THE EXTRACTED FEATURES OF DIFFERENT LEVELS. '#' INDICATES THE DIMENSION OF EACH FEATURE.

Levels	Generality	Short Description	#
Low	*Generic*	GIST features [63] [64]	512
	Special	Color and texture [11]	48
Mid	*Generic*	Attributes [64] [23]	102
	Special	Principles-of-art [2]	165
High	*Generic*	Adjective noun pairs [1]	1200
	Special	Facial expressions [30]	8

The extracted features are abbreviated as GIST, Elements, Attributes, Principles, ANP and Expressions with dimension 512, 48, 102, 165, 1200 and 8, respectively, as summarized in Table III. Please refer to [26] for details.

V. Emotion Distribution Prediction Algorithms

The emotion distribution prediction task of Eq. (2) can be viewed as a regression problem. We detail the proposed MTSSR method together with several baseline algorithms.

A. Baseline A: Global Weighting

The idea of global weighting (GW) is simple and direct. The emotion distribution parameters $\boldsymbol{\theta}_n (n = 1, \cdots, N)$ of all training images are considered as basis functions. The test distribution parameter $\widehat{\boldsymbol{\vartheta}}_m$ is computed by weighting all the basis functions as follows

$$\widehat{\boldsymbol{\vartheta}}_m = \frac{\sum_{n=1}^{N} s_n \boldsymbol{\theta}_n}{\sum_{n=1}^{N} s_n}, \tag{3}$$

where $s_n = \exp(-d(\mathbf{g}_m, \mathbf{f}_n)/\sigma)$ is the similarity between images $\mathbf{g}_m$ and $\mathbf{f}_n$, $d(\cdot,\cdot)$ is a specified distance function, while σ is set as the average distance of all the training images. In experiment, the Euclidean distance is used for $d(\cdot,\cdot)$ and each $\boldsymbol{\theta}_n$, $\widehat{\boldsymbol{\vartheta}}_m$ is reshaped as a column vector for convenience.

B. Baseline B: K-Nearest Neighbor Weighting

Different from GW, K-nearest neighbor weighting (KNNW) just weighs K instead of all basis functions by selecting the top K most similar training images. Suppose the top K largest similarities in $[s_1, \cdots, s_N]$ are $s_{t_1}, \cdots, s_{t_K}$,

then the test parameter $\widehat{\boldsymbol{\vartheta}}_m$ estimated by K-nearest neighbor weighting is computed by

$$\widehat{\boldsymbol{\vartheta}}_m = \frac{\sum_{k=1}^{K} s_{t_k} \boldsymbol{\theta}_{t_k}}{\sum_{k=1}^{K} s_{t_k}}. \tag{4}$$

When $K == N$, KNNW turns to GW.

C. Baseline C: Support Vector Regression

Support vector regression (SVR) aims to find support vectors which lie on the maximum margin hyperplanes in feature space and contribute to predictions. Training SVR means solving

$$\min \frac{1}{2}\|\boldsymbol{w}_i\|_2, s.t. \begin{cases} \theta_{ni} - \langle \boldsymbol{w}_i, \mathbf{f}_n \rangle - b \le \epsilon, \\ \langle \boldsymbol{w}_i, \mathbf{f}_n \rangle + b - \theta_{ni} \le \epsilon, \end{cases} \tag{5}$$

where the target value θ_{ni} is the ith component of $\boldsymbol{\theta}_n$ ($n = 1, \cdots, N$), the inner product plus intercept $\langle \boldsymbol{w}_i, \mathbf{f}_n \rangle + b$ is the prediction for that sample, and ϵ is a free parameter that serves as a threshold. After optimization, we can predict $\widehat{\vartheta}_{mi}$ by $\widehat{\vartheta}_{mi} = \langle \boldsymbol{w}_i, \mathbf{g}_m \rangle + b$. We use the LIBSVM toolbox[5] with linear kernel (for fast speed) to implement SVR for emotion distribution prediction.

D. Algorithm D: Shared Sparse Regression

Suppose $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_N]$, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N]$. The basic idea of shared sparse regression (SSR) is that $\mathbf{g}_m$ and $\widehat{\boldsymbol{\vartheta}}_m$ can be written in terms of bases $\mathbf{F}$ and $\boldsymbol{\Theta}$ respectively, but with shared sparse coefficients $\boldsymbol{\phi}_m$. That is

$$\mathbf{g}_m = \mathbf{F}\boldsymbol{\phi}_m \quad \text{and} \quad \widehat{\boldsymbol{\vartheta}}_m = \boldsymbol{\Theta}\boldsymbol{\phi}_m, \tag{6}$$

where $\boldsymbol{\phi}_m$ is obtained by

$$\begin{aligned} \boldsymbol{\phi}_m^* = & \underset{\boldsymbol{\phi}_m}{\operatorname{argmin}} \|\mathbf{F}\boldsymbol{\phi}_m - \mathbf{g}_m\|^2 + \eta\|\boldsymbol{\phi}_m\|_0, \\ & \text{s.t. } \boldsymbol{\phi}_m \ge 0 \text{ and } \|\boldsymbol{\phi}_m\|_1 = 1, \end{aligned} \tag{7}$$

where η is a regularization coefficient that controls the relative importance of the regularization term and the sum-of-squares error term. The constraints $\boldsymbol{\phi}_m \ge 0$ and $\|\boldsymbol{\phi}_m\|_1 = 1$ ensure that the predicted mixing coefficients in Eq. (1) sum to 1 and that the covariance matrixes are positive definite. In practice, η is decided by cross validation.

By iteratively reweighted least squares (IRLS) [69], [70], the objective function of Eq. (7) can be reduced to the following quadratic function with respect to $\boldsymbol{\phi}_m$

$$\begin{aligned} \mathcal{J}(\boldsymbol{\phi}_m) &\simeq \|\mathbf{F}\boldsymbol{\phi}_m - \mathbf{g}_m\|^2 + \eta \sum_{n=1}^{N} |\phi_{m,n}|_p^p \ (0 \le p \le 1) \\ &\simeq \|\mathbf{F}\boldsymbol{\phi}_m - \mathbf{g}_m\|^2 + \eta \sum_{n=1}^{N} \frac{1}{|\phi_{m,n}|^{2-p} + \varepsilon} |\phi_n|^2 \\ &= \boldsymbol{\phi}_m^{\mathrm{T}}(\mathbf{F}^{\mathrm{T}}\mathbf{F} + \eta\boldsymbol{\Gamma}_m)\boldsymbol{\phi}_m - 2\mathbf{g}_m^{\mathrm{T}}\mathbf{F}\boldsymbol{\phi}_m, \end{aligned} \tag{8}$$

where $\varepsilon > 0$ is introduced to avoid division by zero, $\boldsymbol{\Gamma}_m$ is a diagonal matrix with $\boldsymbol{\Gamma}_m(n,n) = \frac{1}{|\phi_{m,n}|^{2-p} + \varepsilon}$. The optimization problem Eq. (8) can now be easily solved by off-the-shelf optimization methods. In experiment, $p \to 0$. The learning procedure is summarized in Algorithm 1.

[5]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Algorithm 1: Learning procedure for Shared Sparse Regression

Input: Training examples $(\mathbf{F}, \boldsymbol{\Theta})$, test image $\mathbf{g}_m$, error threshold γ, regularization coefficient η, max-epochs E, stability parameter ε

Output: Predicted emotion distribution parameter $\widehat{\boldsymbol{\vartheta}}_m$ for $\mathbf{g}_m$

```
Initialize φ_m^(0);
for e ← 1 to E do
    Γ_m^(e)(n,n) ← 1 / (|φ_{m,n}^(e-1)|^2 + ε), n = 1, ..., N;
    φ_m^(e) ← argmin φ_m^(e-1)T (F^T F + η Γ_m^(e)) φ_m − 2 g_m^T F φ_m^(e-1);
    if ||φ_m^(e) − φ_m^(e-1)||_2 < γ then
        break;
    end
end
return ϑ̂_m = Θ φ_m^(e).
```

E. Algorithm E: Multi-Task Shared Sparse Regression

GM, KNNW and SSR model one test image each time, while SVR predicts one target value each time. They do not explore the latent correlation between different prediction tasks by jointly combining them together. That is, they ignore the task relatedness. Multi-task shared sparse regression (MTSSR) utilizes this information. Different related tasks are learnt simultaneously by extracting and utilizing appropriate shared information across tasks. Compared with SSR, MTSSR argues that the regression performance can be improved by taking advantage of the feature group structure [60].

Suppose $\mathbf{G} = [\mathbf{g}_1, \cdots, \mathbf{g}_M]$, $\boldsymbol{\Omega} = [\widehat{\boldsymbol{\vartheta}}_1, \cdots, \widehat{\boldsymbol{\vartheta}}_M]$, MTSSR jointly predicts $\boldsymbol{\Omega}$ for $\mathbf{G}$ by letting the test features and the target values share the same coefficients $\boldsymbol{\Phi}$ on training data $\mathbf{F}$ and $\boldsymbol{\Theta}$ as follows

$$\mathbf{G} = \mathbf{F}\boldsymbol{\Phi} \quad \text{and} \quad \boldsymbol{\Omega} = \boldsymbol{\Theta}\boldsymbol{\Phi}. \tag{9}$$

$\boldsymbol{\Phi} \in \mathbb{R}^{N \times M}$ is obtained by solving the following convex optimization problem

$$\begin{aligned} & \min_{\Phi} \ \|\mathbf{F}\boldsymbol{\Phi} - \mathbf{G}\|^2 + \eta_1\|\boldsymbol{\Phi}\|_0 + \eta_2\|\boldsymbol{\Phi}\|_{2,1}, \\ & \text{s.t. } \boldsymbol{\phi}_m \ge 0 \text{ and } \|\boldsymbol{\phi}_m\|_1 = 1, \text{ for } m = 1, 2, \cdots, M, \end{aligned} \tag{10}$$

where η_1 and η_2 are regularization coefficients, similar to η in Eq. (7), while $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$-norm of a matrix $\|\boldsymbol{\Phi}\|_{2,1} = \sum_{n=1}^{N} \sqrt{\sum_{m=1}^{M} \phi_{n,m}^2}$. The constraints $\boldsymbol{\phi}_m \ge 0$ and $\|\boldsymbol{\phi}_m\|_1 = 1$ for each m ensure that for each test image, the predicted mixing coefficients in Eq. (1) sum to 1 and that the covariance matrixes are positive definite.

Sparse multi-task regression was previously proposed in [60] for brain imaging identification, which aims to optimize

$$\min_{\Phi} \ \|\mathbf{F}\boldsymbol{\Phi} - \mathbf{G}\|^2 + \eta_1\|\boldsymbol{\Phi}\|_1 + \eta_2\|\boldsymbol{\Phi}\|_{2,1}. \tag{11}$$

The difference between Eq. (10) and Eq. (11) is that the proposed MTSSR utilizes ℓ_0-norm instead of ℓ_1-norm and is optimized with constraints. Please note that though mathematically similar, Eq. (11) is not suitable for continuous emotion distribution prediction, since it cannot guarantee the predicted covariance matrix is positive definite. In such cases, the solution of Eq. (11) violates the basic bidimensional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2617741, IEEE Transactions on Multimedia

Algorithm 2: Learning procedure for Multi-Task Shared Sparse Regression

Input: Training examples $(\mathbf{F}, \mathbf{\Theta})$, test images $\mathbf{G}$, error threshold γ, regularization coefficient η_1, η_2, max-epochs E, stability parameter ε

Output: Predicted emotion distribution parameters $\mathbf{\Omega}$ for $\mathbf{G}$

Initialize $\mathbf{\Phi}^{(0)}$;

for $e \leftarrow 1$ *to* E **do**

  $\varphi_{n,m}^{(e)} = \dfrac{1}{|\phi_{n,m}^{(e)}|^2 + \varepsilon}$, $\psi_n^{(e)} = \dfrac{1}{\sqrt{\sum_m \phi_{n,m}^{(e)2} + \varepsilon}}$,

  $\mathbf{W}_m^{(e)}(n,n) = \eta_1 \varphi_{n,m} + \eta_2 \psi_n$;

  $\mathbf{\Phi}^{(e)} \leftarrow$

  $\operatorname{argmin} \sum_m \|\mathbf{F}\boldsymbol{\phi}_m^{(e-1)} - \mathbf{g}_m\|^2 + \sum_m \boldsymbol{\phi}_m^{(e-1)\mathrm{T}} \mathbf{W}_m^{(e)} \boldsymbol{\phi}_m^{(e-1)}$;

  if $\|\boldsymbol{\phi}_m^{(e)} - \boldsymbol{\phi}_m^{(e-1)}\|_2 < \gamma$ *for each* ϕ_m **then**

    break;

  end

end

return $\mathbf{\Omega} = \mathbf{\Theta}\mathbf{\Phi}$.

Gaussian distribution assumption as in Section III-B.

We employ the iteratively reweighted least squares [69], [70] to optimize Eq. (10). The components of Eq. (10) are transformed by

$$\|\mathbf{\Phi}\|_0 \simeq \sum\nolimits_{n,m} |\phi_{n,m}|_p^p \simeq \sum\nolimits_{n,m} \frac{\phi_{n,m}^2}{|\phi_{n,m}|^{2-p} + \varepsilon}, \quad (12)$$

$$\|\mathbf{\Phi}\|_{2,1} = \sum\nolimits_n \sqrt{\sum\nolimits_m \phi_{n,m}^2} \simeq \sum\nolimits_n \left(\frac{\sum_m \phi_{n,m}^2}{\sqrt{\sum_m \phi_{n,m}^2 + \varepsilon}} \right), \quad (13)$$

where $0 \le p \le 1$. Let $\varphi_{n,m} = 1/(|\phi_{n,m}|^{2-p} + \varepsilon)$ and $\psi_n = 1/\sqrt{\sum_m \phi_{n,m}^2 + \varepsilon}$, then the objective function of Eq. (10) is transformed to

$$\begin{aligned} \mathcal{J}(\mathbf{\Phi}) &\simeq \sum\nolimits_m \|\mathbf{F}\boldsymbol{\phi}_m - \mathbf{g}_m\|^2 + \sum\nolimits_{n,m} (\eta_1 \varphi_{n,m} + \eta_2 \psi_n) \phi_{n,m}^2 \\ &= \sum\nolimits_m \|\mathbf{F}\boldsymbol{\phi}_m - \mathbf{g}_m\|^2 + \sum\nolimits_m \boldsymbol{\phi}_m^{\mathrm{T}} \mathbf{W}_m \boldsymbol{\phi}_m, \end{aligned} \quad (14)$$

where $\mathbf{W}_m$ is a diagonal matrix with $\mathbf{W}_m(n,n) = \eta_1 \varphi_{n,m} + \eta_2 \psi_n$. $\min \mathcal{J}(\mathbf{\Phi})$ is a quadratic programming problem, which can be easily solved by off-the-shelf optimization methods. In experiment, $p \to 0$. The learning procedure is summarized in Algorithm 2.

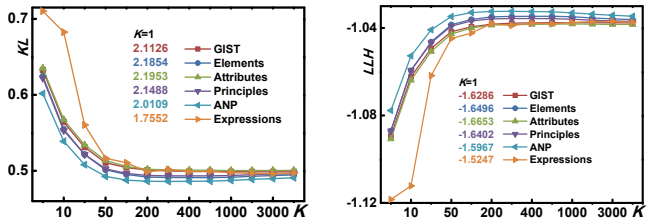
VI. Experiments

To evaluate the effectiveness of the proposed method for continuous distribution prediction of dimensional image emotions, we carried out experiments on 18,700 images which are selected from the Image-Emotion-Social-Net dataset.

A. Evaluation Criteria

The Kullback-Leibler divergence and the log likelihood (abbreviated as KL and LLH) are used as the evaluation metric. As a classical measure of distance between distributions, the KL divergence of the predicted distribution $\widehat{q}_m(\mathbf{x}; \widehat{\boldsymbol{\vartheta}}_m)$ from the ground truth distribution $q_m(\mathbf{x}; \boldsymbol{\vartheta}_m)$ is defined as

$$KL(q_m || \widehat{q}_m) = -\int q_m(\mathbf{x}; \boldsymbol{\vartheta}_m) \ln \left\{ \frac{\widehat{q}_m(\mathbf{x}; \widehat{\boldsymbol{\vartheta}}_m)}{q_m(\mathbf{x}; \boldsymbol{\vartheta}_m)} \right\} d\mathbf{x}. \quad (15)$$



(a) KL divergence (b) Log likelihood

Fig. 8. The influence of K in KNNW on continuous emotion distribution prediction using different features.

In practice, $KL(q_m || \widehat{q}_m)$ is approximated by a finite sum of the points $\{\mathbf{s}_1, \cdots, \mathbf{s}_S\}$ sampled following distribution $q_m(\mathbf{x}; \boldsymbol{\vartheta}_m)$ by

$$KL(q_m || \widehat{q}_m) \simeq \frac{1}{S} \sum_{n=1}^{S} \left\{ \ln q_m(\mathbf{s}_n; \boldsymbol{\vartheta}_m) - \ln \widehat{q}_m(\mathbf{s}_n; \widehat{\boldsymbol{\vartheta}}_m) \right\}. \quad (16)$$

KL measures the amount of information lost when $\widehat{q}_m(\mathbf{x}; \widehat{\boldsymbol{\vartheta}}_m)$ is used to approximate $q_m(\mathbf{x}; \boldsymbol{\vartheta}_m)$. Its value is equal to the expected number of extra bits required to code samples from $q_m(\mathbf{x}; \boldsymbol{\vartheta}_m)$ using a code optimized for $\widehat{q}_m(\mathbf{x}; \widehat{\boldsymbol{\vartheta}}_m)$ rather than the code optimized for $q_m(\mathbf{x}; \boldsymbol{\vartheta}_m)$[6]. $KL \ge 0$ and lower value indicates better performance, with equality if, and only if the predicted distribution $\widehat{q}_m(\mathbf{x}; \widehat{\boldsymbol{\vartheta}}_m)$ is equal to the ground truth distribution $q_m(\mathbf{x}; \boldsymbol{\vartheta}_m)$.

The log likelihood metric is computed based on the actual VA labels $\{\mathbf{x}_{m1}, \cdots, \mathbf{x}_{mR_m}\}$ by

$$LLH(m) = \frac{1}{R_m} \log \prod_{r=1}^{R_m} \widehat{q}_m(\mathbf{x}_{mr}; \widehat{\boldsymbol{\vartheta}}_m) = \frac{1}{R_m} \sum_{r=1}^{R_m} \log \widehat{q}_m(\mathbf{x}_{mr}; \widehat{\boldsymbol{\vartheta}}_m). \quad (17)$$

Higher LLH represents better performance, indicting that the predicted distribution can more accurately fit the actual labels[7].

In experiments, we employ 5-fold (noted as A to E) cross validation [2], [11], [13]. Each run, one fold is selected for testing and the other four folds are used for training. The parameters in our method are selected from the training data. For example, we first select A as the test set. Then we split the data in B to E into 5 folds again, and a new 5-fold cross validation is conducted to select the best parameters based on the average KL. The selected parameters are used to test A. We computed the average KL, LLH and the standard deviation of the 5 runs. To better explain the validation process, we also reported the performances of 1 run.

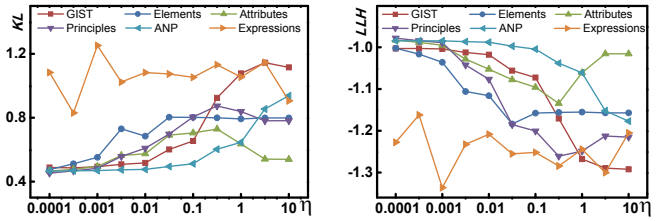
B. Results and Discussions

1) On the Influence of K in KNNW: Firstly, we investigated the influence of K in KNNW on the performance of emotion distribution prediction (K = 1, 5, 10, 20, 50, 100, 200, 300, 400, 500, 1000, 2000, 3000, 4000). When $K = 1$, KNNW refers to nearest neighbor weighting (NNW). The results are illustrated in Figure 8. The results of NNW are given in numerical values for better illustration. It is clear to see that (1) for each feature, NNW performs worst, meaning

[6]https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

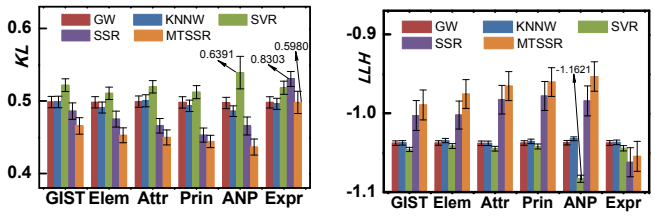
[7]https://en.wikipedia.org/wiki/Likelihood_function

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2617741, IEEE Transactions on Multimedia



(a) KL divergence (b) Log likelihood

Fig. 9. The influence of η in SSR on continuous emotion distribution prediction using different features.



(a) KL divergence (b) Log likelihood

Fig. 12. Performance comparison between the proposed method and the three baselines on emotion distribution prediction using different features.

TABLE IV
STATISTICAL SIGNIFICANCE TEST OF MTSSR COMPARED WITH THE THREE BASELINES AND SSR MEASURED BY P-VALUE ($\times 10^{-3}$).

	KL divergence				Log likelihood			
	GW	KNNW	SVR	SSR	GW	KNNW	SVR	SSR
GIST	2.80	2.79	1.16	17.95	16.28	16.64	9.55	8.22
Elem	1.76	2.35	1.07	0.45	12.97	15.84	9.96	0.12
Attr	2.58	2.51	1.09	1.11	19.98	20.27	13.00	4.00
Prin	2.17	2.78	1.14	1.90	11.35	12.87	8.73	0.42
ANP	2.90	5.09	0.20	0.13	18.41	25.00	1.30	0.01
Expr	1.71	1.79	0.72	5.33	9.76	10.45	6.74	14.65

that using only one training image with the most similar features to the test image to predict the emotion distribution is insufficient; (2) the best K is dependent on the extracted features and is almost consistent on KL and LLH for each feature; (3) $K = 4000$, 500, 1000, 400, 300 and 3000 perform best for features GIST, Elements, Attributes, Principles, ANP and Expressions, respectively. These best Ks are selected as baselines for comparison with the proposed (MT)SSR.

2) On the Influence of η in SSR: Secondly, we evaluated the influence of the regularization parameter η in the proposed SSR on emotion distribution prediction (η = 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10). The results of the average KL divergence and log likelihood are shown in Figure 9. Generally, with the decrease of η, the performance becomes better. When η decreases to $O(10^{-4})$, the performance turns to be stable. $\eta = 0.0005$, 0.0001, 0.0001, 0.0001, 0.0001 and 0.0005 perform best for features GIST, Elements, Attributes, Principles, ANP and Expressions, respectively.

3) On the Influence of η_1, η_2 in MTSSR: Finally, the influences of the regularization parameters η_1, η_2 in the proposed MTSSR are validated, with results shown in Figure 10 and Figure 11. For clarity, η_1 = 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5 are plotted with η_2 = 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100(1000). From these results, we can find that (1) generally, with the decrease of η_1, the performance becomes better; when η_1 decreases to $O(10^{-4})$, the performance turns to be stable, which is similar to SSR; (2) for each η_1, with the increase of η_2, the performance firstly becomes better and then turns worse, meaning that there exists the best η_2. So we can conclude that selecting proper η_2 can indeed improve the performance of emotion distribution prediction, which indicates the significance of the multi-task learning settings.

4) On Different Methods and Features: We compared the performance of the proposed method with the three baselines on different features. The average KL and LLH and the standard deviation are illustrated in Figure 12, while the statistical significance test is shown in Table IV.

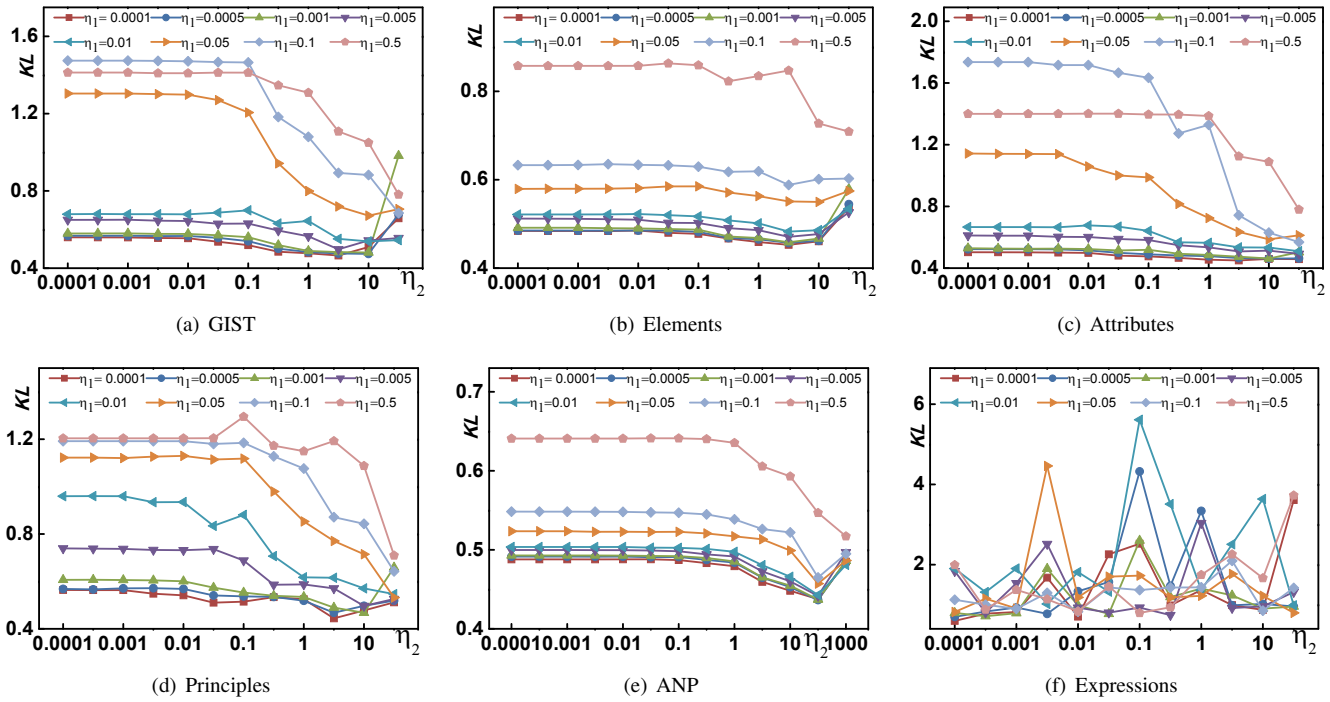
From these results, we can find that (1) KL and LLH are dependent on both the features and the models; they are relatively consistent to measure the performance of distribution prediction; (2) for all the features except Expressions, the proposed MTSSR model significantly outperforms the three baselines and SSR under 95% confidence interval, which demonstrates the effectiveness of MTSSR in emotion distribution prediction; (3) MTSSR outperforms SSR with an performance improvement of 4.4%, 4.9%, 3.6%, 2.0%, 6.7% on KL and 38.8% on KL and 1.4%, 2.7%, 1.8%, 1.8%, 3.2% and 10.2% on LLH for the six kinds of features respectively; this superiority benefits from the exploration of latent information between different tasks; (4) the best features are ANP, ANP, Elements, Principles and ANP for GW, KNNW, SVR, SSR and MTSSR, respectively; Generally, the low-level generic features perform the worst, which indicates that they cannot represent image emotions well because of the largest "affective gap"; More interpretable mid-level and the high-level features have stronger link to image emotions, which is consistent with the conclusions in [26]; (5) though simple, GW and KNNW outperform SVR on average in emotion distribution prediction; (6) the best KL divergence of all the methods are still larger than 0,4, indicating that the emotion distribution prediction is a challenging task and that current methods still cannot model this task accurately.

Using related best features, we show some detailed prediction results of different methods in Figure 6(d) to Figure 6(h). Though not very obvious from the contour lines, the predicted distributions of the proposed (MT)SSR are more similar to the ground truth distribution than the three baselines, which can be clearly observed by comparing the values of the distribution parameters. The predicted results of all methods, especially GW and KNNW, tend to be close to the average values of the parameters, caused by the assumption that the test parameters can be linearly represented by the training parameters with positive coefficients, which ensures the positive definiteness of covariance matrixes. But in such cases, the smallest or largest parameters cannot be well predicted, since they cannot be well linearly represented by the training parameters.

VII. CONCLUSION

In this paper, we proposed to predict the continuous probability distribution of image emotions represented in VA space, which can be viewed as an initial attempt to measure the subjective evaluation of human emotion perceptions. We presented

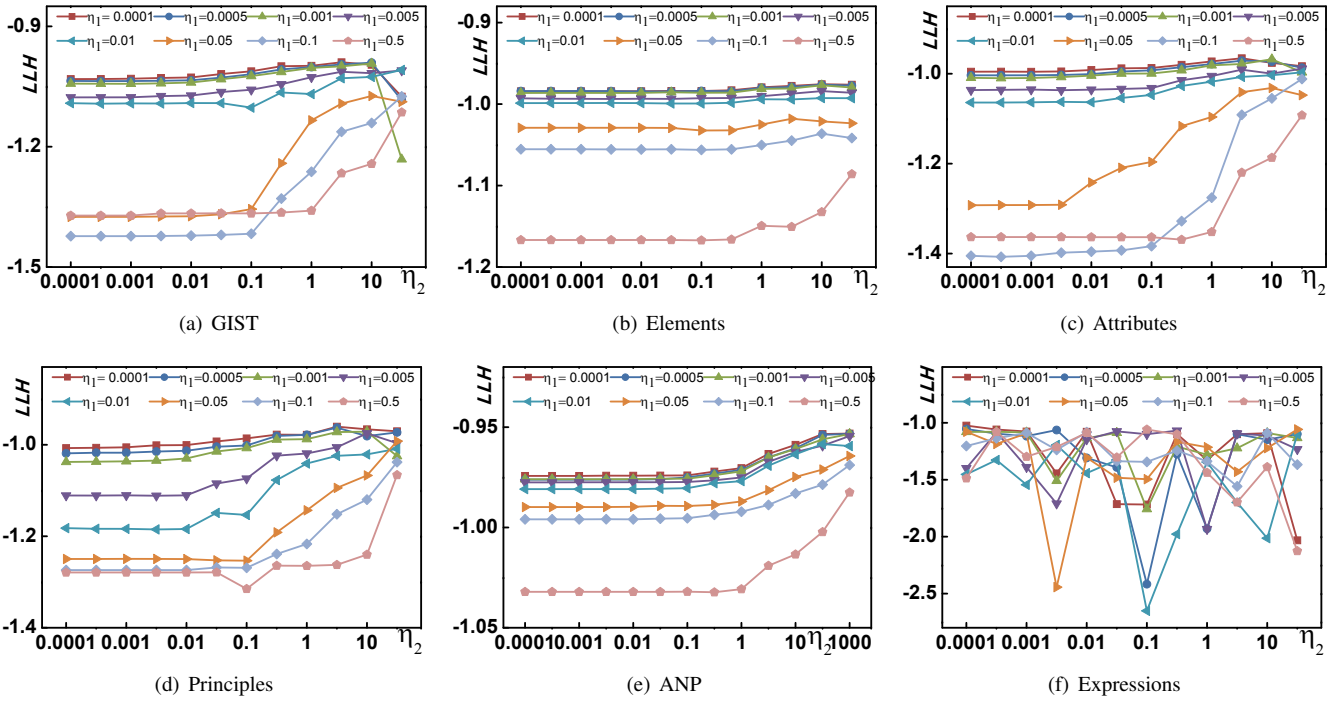
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2016.2617741, IEEE Transactions on Multimedia



(a) GIST (b) Elements (c) Attributes

(d) Principles (e) ANP (f) Expressions

Fig. 10. The influence of η_1, η_2 in MTSSR on continuous emotion distribution prediction using different features on KL divergence.



(a) GIST (b) Elements (c) Attributes

(d) Principles (e) ANP (f) Expressions

Fig. 11. The influence of η_1, η_2 in MTSSR on continuous emotion distribution prediction using different features on log likelihood.

(multi-task) shared sparse regression as the learning model and optimized it by iteratively reweighted least squares. Besides, different levels of emotion features were extracted and three baseline algorithms were provided. Experiments conducted on the Image-Emotion-Social-Net dataset corroborated the effectiveness of the proposed method. The predicted emotion distribution can be explored in many applications, such as affective image retrieval and emotion transfer.

For further studies, we will consider exploring social related factors [17], [71], such as social correlations, known locations and personal interests, for emotion distribution prediction. Consistently combining and fusing multi-modal features [72] in MTSSR may further improve the prediction performance, which is also worth studying. In addition, we will try deep learning for emotion distribution prediction.

REFERENCES

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM International Conference on Multimedia*, 2013, pp. 223–232.

[2] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *ACM International Conference on Multimedia*, 2014, pp. 47–56.

[3] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *ACM International Conference on Multimedia*, 2014, pp. 367–376.

[4] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[5] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[6] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 860–868.

[7] S. Zhao, H. Yao, X. Jiang, and X. Sun, "Predicting discrete probability distribution of image emotions," in *IEEE International Conference on Image Processing*, 2015, pp. 2459–2463.

[8] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T.-S. Chua, "Predicting personalized emotion perceptions of social images," in *ACM International Conference on Multimedia*, 2016.

[9] S. Zhao, H. Yao, and X. Jiang, "Predicting continuous probability distribution of image emotions in valence-arousal space," in *ACM International Conference on Multimedia*, 2015, pp. 879–882.

[10] W.-n. Wang, Y.-l. Yu, and S.-m. Jiang, "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.

[11] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM International Conference on Multimedia*, 2010, pp. 83–92.

[12] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *IEEE International Conference on Image Processing*, 2008, pp. 101–104.

[13] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *ACM International Conference on Multimedia*, 2012, pp. 229–238.

[14] B. Li, W. Xiong, W. Hu, and X. Ding, "Context-aware affective images classification based on bilayer sparse representation," in *ACM International Conference on Multimedia*, 2012, pp. 721–724.

[15] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. C. M. Fong, "Quantitative study of individual emotional states in social networks," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 132–144, 2012.

[16] 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 *ACM International Conference on Multimedia*, 2012, pp. 857–860.

[17] 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 *AAAI Conference on Artificial Intelligence*, 2014, pp. 306–312.

[18] H. Schlosberg, "Three dimensions of emotion." *Psychological Review*, vol. 61, no. 2, p. 81, 1954.

[19] S. Benini, L. Canini, and R. Leonardi, "A connotative space for supporting movie affective recommendation," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1356–1370, 2011.

[20] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[21] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

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

[23] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *ACM International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2013, p. 10.

[24] Y. Yang, P. Cui, W. Zhu, and S. Yang, "User interest and social influence based emotion prediction for individuals," in *ACM International Conference on Multimedia*, 2013, pp. 785–788.

[25] Y. Yang, P. Cui, W. Zhu, H. V. Zhao, Y. Shi, and S. Yang, "Emotionally representative image discovery for social events," in *ACM International Conference on Multimedia Retrieval*, 2014, p. 177.

[26] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *ACM International Conference on Multimedia*, 2014, pp. 1025–1028.

[27] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2559–2566.

[28] S. Zheng, A. Yuille, and Z. Tu, "Detecting object boundaries using low-, mid-, and high-level information," *Computer Vision and Image Understanding*, vol. 114, no. 10, pp. 1055–1067, 2010.

[29] Y. Guo, G. Ding, X. Jin, and J. Wang, "Learning predictable and discriminative attributes for visual recognition." in *AAAI Conference on Artificial Intelligence*, 2015, pp. 3783–3789.

[30] P. Yang, Q. Liu, and D. N. Metaxas, "Exploring facial expressions with compositional features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2638–2644.

[31] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji, "Video indexing and recommendation based on affective analysis of viewers," in *ACM International Conference on Multimedia*, 2011, pp. 1473–1476.

[32] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[33] 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.

[34] S. Zhao, H. Yao, and X. Sun, "Video classification and recommendation based on affective analysis of viewers," *Neurocomputing*, vol. 119, pp. 101–110, 2013.

[35] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[36] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2015.

[37] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic, "Affective labeling in a content-based recommender system for images," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 391–400, 2013.

[38] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[39] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[40] 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.

[41] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, p. 40, 2012.

[42] Y.-H. Yang and J.-Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1304–1315, 2013.

[43] P.-C. Chen, K.-S. Lin, and H. H. Chen, "Emotional accompaniment generation system based on harmonic progression," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1469–1479, 2013.

[44] S. Zhao, H. Yao, F. Wang, X. Jiang, and W. Zhang, "Emotion based image musicalization," in *IEEE International Conference on Multimedia and Expo Workshops*, 2014, pp. 1–6.

[45] S. Zhao, H. Yao, X. Sun, X. Jiang, and P. Xu, "Flexible presentation of videos based on affective content analysis," in *International Conference on Multimedia Modelling*, 2013, pp. 368–379.

[46] K. Yadati, H. Katti, and M. Kankanhalli, "Cavva: Computational affective video-in-video advertising," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 15–23, 2014.

[47] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.

[48] S. E. Shepstone, Z.-H. Tan, and S. H. Jensen, "Using audio-derived affective offset to enhance tv recommendation," *IEEE Transactions on Multimedia*, vol. 16, no. 7, pp. 1999–2010, 2014.

[49] S. Wang and Q. Ji, "Video affective content analysis: a survey of state of the art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.

[50] M. Carney, P. Cunningham, J. Dowling, and C. Lee, "Predicting probability distributions for surf height using an ensemble of mixture density networks," in *ACM International Conference on Machine Learning*, 2005, pp. 113–120.

[51] H. Liu, Z. Hu, D. Zhou, and H. Tian, "Cumulative probability distribution model for evaluating user behavior prediction algorithms," in *IEEE International Conference on Social Computing*, 2013, pp. 385–390.

[52] G. Pipa, S. Grün, and C. van Vreeswijk, "Impact of spike train autostructure on probability distribution of joint spike events," *Neural Computation*, vol. 25, no. 5, pp. 1123–1163, 2013.

[53] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[54] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[55] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," *Arizona State University*, 2012.

[56] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer, "Multi-task learning for hiv therapy screening," in *ACM International Conference on Machine Learning*, 2008, pp. 56–63.

[57] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," *Advances in Neural Information Processing Systems*, vol. 19, p. 41, 2007.

[58] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 895–903.

[59] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *The Journal of Machine Learning Research*, vol. 6, pp. 1817–1853, 2005.

[60] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin, L. Shen *et al.*, "Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance," in *IEEE International Conference on Computer Vision*, 2011, pp. 557–562.

[61] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[62] 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.

[63] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[64] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2751–2758.

[65] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 771–778.

[66] H. Zhang, Y. Yang, H. Luan, S. Yang, and T.-S. Chua, "Start from scratch: Towards automatically identifying, modeling, and naming visual attributes," in *ACM International Conference on Multimedia*, 2014, pp. 187–196.

[67] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.

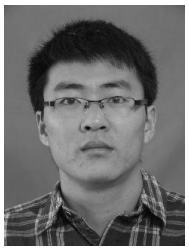
[68] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[69] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3869–3872.

[70] C. Chen, J. Huang, L. He, and H. Li, "Preconditioning for accelerated iteratively reweighted least squares in structured sparsity reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2713–2720.

[71] Y. Gao, S. Zhao, Y. Yang, and T.-S. Chua, "Multimedia social event detection in microblog," in *International Conference on Multimedia Modeling*, 2015, pp. 269–281.

[72] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2075–2082.



Sicheng Zhao received the Ph.D. degree from Harbin Institute of Technology in 2016. He is now a postdoctoral research fellow in the School of Software, Tsinghua University, China. His research interests include affective computing, social media analysis and multimedia information retrieval.

Hongxun Yao received the B.S. and M.S. degrees from Harbin Shipbuilding Engineering Institute, China, in 1987 and 1990, respectively, and the Ph.D. degree from Harbin Institute of Technology, China, in 2003. She is a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. She has authored six books and has published over 200 scientific papers. Her research interests include computer vision, pattern recognition, multimedia computing, and human–computer interaction technology.

Yue Gao received the B.S. degree from the Harbin Institute of Technology, Harbin, China, and the M.E. and Ph.D. degrees from Tsinghua University, Beijing, China.

Rongrong Ji is a Professor, the Director of the Intelligent Multimedia Technology Laboratory, and the Dean Assistant of the School of Information Science and Engineering with Xiamen University, Xiamen, China. He has authored over 100 paper published in international journals and conferences. His research interests include innovative technologies for multimedia signal processing, computer vision, and pattern recognition.

Guiguang Ding received the Ph.D. degree in electronic engineering from Xidian University, China. He is currently an Associate Professor with the School of Software, Tsinghua University. Since 2006, he has been a postdoctoral research fellow with the Department of Automation, Tsinghua University. His current research centers on the area of multimedia information retrieval, computer vision, and machine learning.