Discrete Probability Distribution Prediction of Image Emotions with Shared Sparse Learning

Sicheng Zhao, Guiguang Ding, Yue Gao, *Senior Member, IEEE*, Xin Zhao, Youbao Tang, Jungong Han, Hongxun Yao, and Qingming Huang, *Fellow, IEEE*

Abstract—Computationally modelling the affective content of images has been extensively studied recently because of its wide applications in entertainment, advertisement, and education. Significant progress has been made on designing discriminative features to bridge the affective gap. Assuming that viewers can reach a consensus on the emotion of images, most existing works focused on assigning the dominant emotion category or the average dimension values to an image. However, the image emotions perceived by viewers are subjective by nature with the influence of personal and situational factors. In this paper, we propose a novel machine learning approach that characterizes the categorical image emotions as a discrete probability distribution (DPD). To associate emotion with the visual features extracted from images, we present shared sparse learning to learn the combination coefficients, with which the DPD of an unseen image is predicted by linearly combining the DPDs of the training images. Furthermore, we extend our method to the setup where multi-features are available and learn the optimal weights for each feature to reflect the importance of different features. Extensive experiments are carried out on Abstract, Emotion6 and IESN datasets and the results demonstrate the superiority of the proposed method, as compared to the state-of-the-art approaches.

Index Terms—Emotion distribution, image emotions, shared sparse learning, multi-feature fusion

1 Introduction

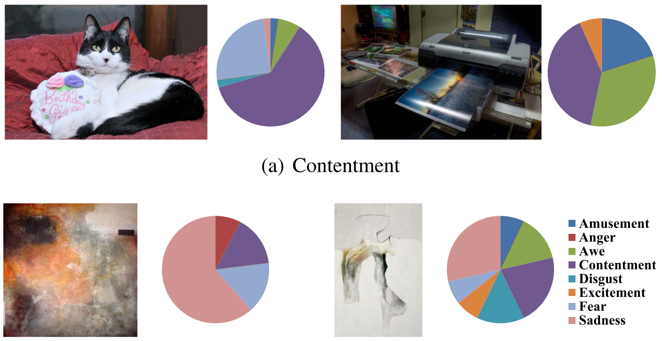
Images play an important role in people's daily lives, which are widely used together with text and videos to share their activities and express their opinions. As the emotions that people perceive from images can usually influence their visual preference and determine their decision making, analyzing images at the emotional level is considered promising for facilitating image understanding (e.g., sentiment concept classification [1]) and human behavior estimation (e.g., stress detection [2]). Driven by such a broad application prospect, scientists have tried to develop computational models to analyze the affective content of images [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. This task is often referred to as image emotion recognition (IER) [6], [10], [14], which can be deemed as a machine learning problem. Typically, IER includes three steps: collecting human annotations of image emotions, extracting visual features from images and employing machine learning techniques to learn the mapping between features and emotions.

Similar to other visual recognition problems, one main challenge for IER is affective gap, which is 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" [19]. In order to bridge the affective gap, effective hand-crafted or learning-based features are designed to express emotions better. Existing IER methods mainly focused on assigning the dominant emotion category or the average dimension values to an image, based on the assumption that viewers can reach a consensus on the emotion of images.

However, labeling the emotions in images is in fact highly inconsistent, which causes the so-called subjective evaluation. That is, viewers might perceive different emotions from the same image due to the influence of various personal and situational factors, such as the cultural background, personality and social context [6], [10], [11], [12], [13], [16], [17]. Fig. 1 illustrates the subjectivity issue for categorical emotions. To train an IER model, the emotion annotations have to be solicited from viewers. The ground-truth annotation of an image is usually obtained using the dominant emotion category, like *Contentment* and *Sadness* in Fig. 1, where the pie chart on the right of each image shows the corresponding emotion distributions. We can see that the two images of each group have the same dominant emotion category but differ a lot in their emotion variances. Therefore, the dominant emotion category does not precisely reflect the affective content of these images.

- *S. Zhao, G. Ding, Y. Gao, and X. Zhao are with the School of Software, Tsinghua University, Beijing 100084, China. E-mail: {schzhao, zhaoxin19}@gmail.com, {dinggg, gaoyue}@tsinghua.edu.cn.*
- *Y. Tang is with the National Institutes of Health, Bethesda, MD 20892. E-mail: tybxiaobao@gmail.com.*
- *J. Han is with the School of Computing & Communications, Lancaster University, Lancaster LA1 4YW, United Kingdom. E-mail: jungonghan77@gmail.com.*
- *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.*
- *Q. Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Huairou 101408, China. E-mail: qmhuang@ucas.ac.cn.*

Manuscript received 27 June 2017; revised 28 Jan. 2018; accepted 12 Mar. 2018. Date of publication 23 Mar. 2018; date of current version 25 Nov. 2020. (Corresponding author: Guiguang Ding.)
Recommended for acceptance by H. Gunes.
Digital Object Identifier no. 10.1109/TAFFC.2018.2818685



(a) Contentment

(b) Sadness

Fig. 1. The differences between affective image classification and emotion distribution prediction. The words (a) *Contentment* and (b) *Sadness* are the target emotion categories to related images by affective image classification, while the pie chart on the right of each image is the target distribution by emotion distribution prediction on 8 emotion categories.

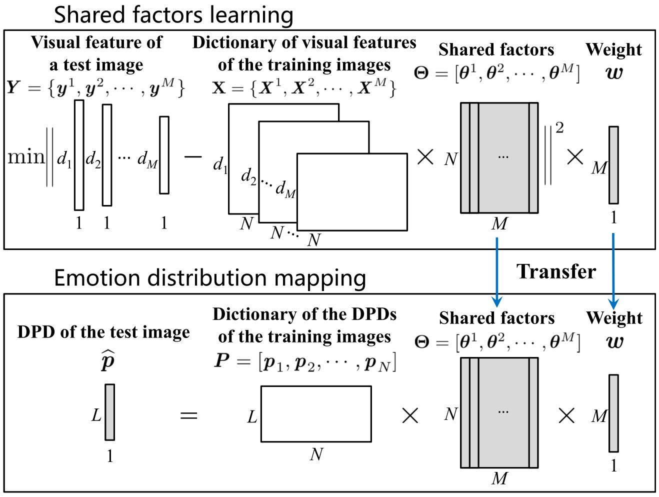

Shared factors learning
Visual feature of a test image
$Y = \{y^1, y^2, \cdots, y^M\}$
Dictionary of visual features of the training images
$\mathbf{X} = \{X^1, X^2, \cdots, X^M\}$
Shared factors
$\Theta = [\theta^1, \theta^2, \cdots, \theta^M]$
Weight
w
min
d_1
d_2
d_M
1
N
M
2
Transfer
Emotion distribution mapping
DPD of the test image
$\widehat{p}$
Dictionary of the DPDs of the training images
$P = [p_1, p_2, \cdots, p_N]$
L

Fig. 2. Diagram of the emotion distribution prediction process. Given the visual feature of a test image, we use a dictionary of visual features of the training images to learn the shared factors, and then use the same shared factors to predict the DPD of the test image by linearly combining the DPDs of the training images. The white and gray boxes are used to denote the observed variables and the variables to be estimated, respectively.

As noted in [20], to tackle the subjectivity challenge, two kinds of IER tasks can be performed: user-centric personalized emotion perception prediction for each viewer [13], [18] and image-centric emotion probability distribution prediction for each image [11], [12], [15]. In this paper, we propose a novel method to predict the DPD of image emotions from visual features, based on the following hypotheses:

- Hypothesis 1: The images, which are close to one another in the visual feature space, would have similar DPDs in the categorical emotion space.
- Hypothesis 2: The DPD of a test image can be approximately modeled as a linear combination of the DPDs of the training images.

Our method mainly involves two processes, as illustrated in Fig. 2. First, it learns a set of combination coefficients (called shared factors) to reconstruct the visual features of a test image with the visual features of the training images. Second, it linearly combines the DPDs of the training images using the learned shared factors to compute the DPD of the test image. The two processes are referred to shared factors learning and emotion distribution mapping, respectively. The shared factors can be learned using various optimization methods with different constraints. Besides four simple baselines, in this paper we formalize the factors learning task as a shared sparse learning (SSL) problem. To fully explore the representation ability of different features, we further present weighted multi-feature shared sparse learning (WMFSSL), which can automatically learn the optimal weights for each feature to reflect their importance. The SSL series problems are optimized by iteratively reweighted least squares (IRLS) [21], [22]. We validate the effectiveness of the proposed method for discrete emotion distribution prediction on Abstract [5], Emotion6 [11] and Image-Emotion-Social-Net (IESN) [13], [18] datasets.

The contributions of this paper are two-fold. First, we present a novel learning model, named shared sparse learning under both uni-feature and multi-feature settings for DPD prediction of image emotions, and optimize it by iteratively reweighted least squares. Second, we provide a systematic summarization on predicting the DPD of image emotions for further research, including the datasets, baselines, and evaluations. To the best of our knowledge, we are the first to employ multi-feature fusion for DPD prediction.

One preliminary conference version on DPD prediction of image emotions was first introduced in our previous work [12], [23]. Our new improvement compared with the conference version lies in the following three aspects: (1) we perform a more comprehensive survey of related works; (2) we provide a more systematic summarization for DPD prediction of image emotions; and (3) we conduct more comparative experiments and enrich the analysis of the results.

The rest of this paper is organized as follows. Section 2 reviews the related works. Section 3 provides an overview of the proposed method. Section 4 presents the emotion distribution prediction algorithms, including the proposed (WMF)SSL and four baseline methods. Section 5 describes the experimental setup, including the datasets, extracted features, evaluation metrics and implementation details. Experimental results and analysis are reported in Section 6, followed by the conclusion in Section 7.

2 Related Works

In this section, we review related works on image emotion recognition, probability distribution prediction, sparse learning and multi-model learning.

Image Emotion Recognition. As an active research topic for several years, IER has attracted some attention from both the academic and industrial communities. Based on the emotion representation models, recognition tasks, extracted features and machine learning methods, we can classify related works into different categories.

There are two typical kinds of emotion representation models: categorical emotion states (CES) and dimensional emotion space (DES). CES methods directly map emotions to one of a few basic categories[1] [3], [4], [5], [7], [8], [14], [26], [27], [28], [29], [30], [31], [32], such as *surprise* and *fear*. DES methods employ 3-D or 2-D space to represent emotions, such as valence-arousal-dominance [33], natural-temporal-energetic [34] and valence-arousal [7], [10], [19].

[1] Specifically, image emotion is often called image sentiment for binary positive or negative classification [1], [9], [24], [25].

The former one is straightforward for users to understand and label, while the latter one is more descriptive. Accordingly, different tasks have been performed, including affective image classification [4], [5], [7], [8], [9], [10], [13], [14], [18], [28], [29], [30], [31], [32], [35], regression [7], [10], [13], [18] and retrieval [3], [36]. As the most popular research task, affective image classification mainly tries to assign a dominant emotion category to an image based on CES models. We also represent image emotions using CES model. But instead of focusing on the single dominant emotion category, we propose to predict the DPD of image emotions.

Feature extraction plays an important role in IER. In the early years, different levels of hand-crafted features are designed to bridge the affective gap. Yanulevskaya et al. [4] extracted low-level holistic *Wiccest* and *Gabor* features. Inspired from psychology and art theory, Machajdik et al. [5] defined a combination of rich features, including *color*, *texture*, *composition* and simple *semantics*. The mapping between emotion and low-level *shape* features is investigated in [7]. Complementary to *elements-of-art*, more robust and invariant visual features are designed based on *principles-of-art* to capture mid-level emotion representation [10]. Visual sentiment ontology and detectors are proposed to detect high-level *adjective noun pairs* based on large-scale social multimedia data [1] [9]. More recently, with the great success of convolutional nerual network (CNN) in many computer vision tasks, such as image classification [37] and object detection [38], CNN has also been directly employed in IER [14], [32], [35]. In this paper, we extract hand-crafted features together with CNN features, and jointly combine them for emotion distribution prediction.

To learn the mapping between features and emotions, different machine learning methods have been employed, such as Naive Bayes [5], SVM or SVR [7], [10], sparse learning [15], [28], multi-graph learning [36], nonlinear matrix completion [35] and hypergraph learning [13], [18]. We present a novel sparse learning method to map the feature space to emotion distribution space.

Note that affective content analysis has also been widely studied based on other types of input data, such as text [39], [40], speech [41], [42], music [43], [44], [45], [46], videos [47], [48], [49], [50], physiological signals [51], [52] and multimodal data [53], [54], [55].

Probability Distribution Prediction. In many applications of machine learning, simply predicting the most likely value for a target variable is not enough. For example, it is often important in economics to study the fluctuations of stocks. In such cases, it would be more reasonable and useful to predict the probability distribution for that variable [56], which has been studied in some areas, such as surf height [56], user behavior [57], spike events [58] and facial ages [59].

According to probability theory, there are typically two types of probability distributions: discrete probability distribution (DPD) and continuous probability distribution (CPD). Generally, the distribution prediction task can be formalized as a regression problem. As the emotions that are evoked in viewers by an image are highly subjective, predicting the distribution instead of the dominant emotion would make more sense. 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 [11], [12]. For DES, the task usually turns to predict the parameters of specified continuous probability distribution, the form of which should be firstly decided, such as Gaussian distribution [15] and exponential distribution. In this paper, we focus on the former one, i.e., predicting the DPD of image emotions.

Sparse Learning and Multi-Modal 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 [60], visual classification [61] and emotion analysis [15], [28]. Meanwhile, in many real-world applications, we might have multi-modal data [62], [63] from different sources [55], [64], [65], [66], [67]. We may also extract multiple features for each modality [36], [68], [69], [70]. As different modal data and different features usually represent different aspects of the target, jointly combining them may promisingly improve the performance [62], [63]. Besides the traditional early fusion and late fusion, there are many other multi-modal/multi-feature fusion strategies, such as hypergraph learning [71], multigrahp learning [72] and multimodal deep learning [73]. By extending sparse learning in multi-feature settings, in this paper we present weighted multi-feature shared sparse learning to make full use of the representation ability of different features for probability distribution prediction of visual emotions.

3 System Overview

Our goal is to predict the DPD of image emotions when multi-features are available. Suppose we have L emotion categories $c_1, c_2, \ldots, c_L$ and N training images $I_1, I_2, \ldots, I_N$. The mth features of the N training images are $\boldsymbol{X}^m = [\boldsymbol{x}_1^m, \boldsymbol{x}_2^m, \ldots, \boldsymbol{x}_N^m]$ and the feature dimension is d_m $(m = 1, 2, \ldots, M)$. Let $\boldsymbol{p}_n = [p_{n1}, \ldots, p_{nl}, \ldots, p_{nL}]^{\mathrm{T}}$ denote the emotion distribution of the image I_n, where p_{nl} represents the probability that image I_n conveys emotion c_l $(n = 1, 2, \ldots, N, l = 1, 2, \ldots, L)$. For each image I_n, we have $\sum_{l=1}^{L} p_{nl} = 1$. Suppose I is a test image, its M features are $\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^M$ and the ground-truth distribution is $\boldsymbol{p} = [p_1, p_2, \ldots, p_L]^{\mathrm{T}}$. Let $\mathbf{X} = \{\boldsymbol{X}^1, \boldsymbol{X}^2, \ldots, \boldsymbol{X}^M\}$ and $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^M\}$ denote the feature set of the training images and the test image, respectively. Let $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N]$ denote the training labels of emotion distribution. Then our task is to predict emotion distribution $\widehat{\boldsymbol{p}} = [\widehat{p}_1, \ldots, \widehat{p}_l, \ldots, \widehat{p}_L]^{\mathrm{T}}$, where $\widehat{p}_l = p(c_l|\boldsymbol{Y})$ for test image I based on training examples $(\mathbf{X}, \boldsymbol{P})$. That is, our task aims to find the appropriate mapping

$$f : \{(\mathbf{X}, \boldsymbol{P}), \boldsymbol{Y}\} \rightarrow \widehat{\boldsymbol{p}}. \tag{1}$$

The framework of the proposed method is shown in Fig. 3, which consists of operations in the visual space and the emotion space. In the visual space, we extract multi-features from the images and use algorithms such as PCA for dimension reduction [5], [10]. In the emotion space, the human emotion annotations are normalized to obtain the DPDs for the training images. For a given test image, the shared factors learning algorithms are used to learn the mapping factors in the visual space, which are directly transferred to the emotion space to predict the DPD.

4 Distribution Prediction Algorithms

In this section, we introduce the proposed emotion distribution prediction method in detail. Since few algorithms have

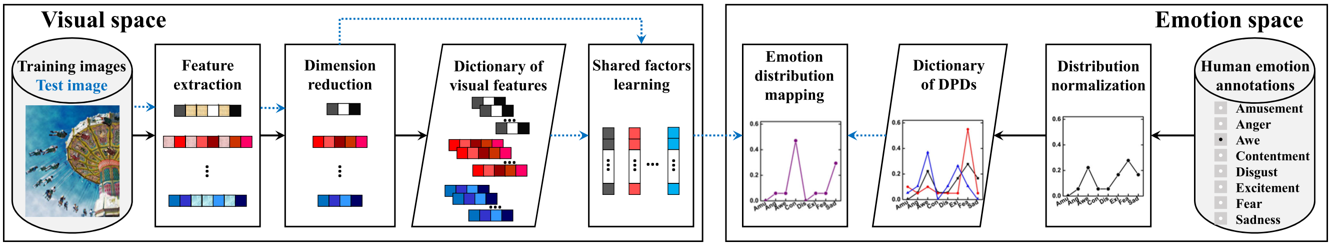


Fig. 3. The framework of the proposed method for DPD prediction of image emotions from visual features. The black solid and blue dash arrowed lines indicate the operations for the training and test images, respectively.

been developed before to predict the DPD of image emotions, we first introduce some baselines and then present our main prediction algorithm. For clarification, we take the mth feature for example to explain the four baselines and SSL.

4.1 Baseline A: Global Weighting

The idea of global weighting (GW) is simple and direct. The emotion distributions $\boldsymbol{p}_n (n = 1, 2, \ldots, N)$ of all training images are considered as basis functions. The emotion distribution $\widehat{\boldsymbol{p}}$ for image I is computed by weighting all the basis functions,

$$\widehat{\boldsymbol{p}} = \frac{\sum_{n=1}^{N} s_n \boldsymbol{p}_n}{\sum_{n=1}^{N} s_n}, \tag{2}$$

where $s_n = \exp(-\frac{d(\boldsymbol{y}^m, \boldsymbol{x}_n^m)}{\sigma})$ is the similarity between $\boldsymbol{y}^m$ and $\boldsymbol{x}_n^m$, $d(\cdot,\cdot)$ is a specified distance function, σ is set as the average distance of all the training images. In our implementation the Euclidean distance is used for $d(\cdot,\cdot)$.

4.2 Baseline B: K-Nearest Neighbor Weighting

Different from GW, K-nearest neighbor weighting (KNNW) just uses K instead of all basis functions by selecting the top K most similar training images. Suppose the top K largest similarities in $[s_1, \ldots, s_N]$ are $s_{l_1}, \ldots, s_{l_K}$, then the emotion distribution $\widehat{\boldsymbol{p}}$ for image I predicted using K-nearest neighbor weighting is computed by

$$\widehat{\boldsymbol{p}} = \frac{\sum_{k=1}^{K} s_{l_k} \mathbf{p}_{l_k}}{\sum_{k=1}^{K} s_{l_k}}. \tag{3}$$

When $K = N$, KNNW equals GW.

4.3 Baseline C: Softmax Regression

Softmax regression (SR) defines the posterior emotion $p(c_l|\boldsymbol{y}^m)$ as a softmax transformation of linear functions of the features $\boldsymbol{y}^m$, which is computed by

$$p(c_l|\boldsymbol{y}^m) = \frac{\exp(\boldsymbol{v}_l^{\mathrm{T}} \boldsymbol{y}^m)}{\sum_{j=1}^{L} \exp(\boldsymbol{v}_j^{\mathrm{T}} \boldsymbol{y}^m)}, \tag{4}$$

where $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L$ are d_m dimensional weight vectors. These parameters $\{\boldsymbol{v}_l\}$ are determined by optimizing the following objective function, which tries to minimize the errors between the predicted results $p(c_j|\boldsymbol{x}_n^m)$ of training examples and their ground truth p_{nj}

$$\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_L^* = \min_{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L} \mathcal{E}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L), \tag{5}$$

$$\mathcal{E}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L) = \sum_{n=1}^{N} \sum_{j=1}^{L} [p(c_j|\boldsymbol{x}_n^m) - p_{nj}]^2 + \lambda \sum_{j=1}^{L} \|\boldsymbol{v}_j\|_2^2, \tag{6}$$

where $\|.\|_2$ is ℓ_2 norm, λ is the regularization coefficient that controls the relative importance of the regularization term and the sum-of-squares error term.

We use gradient descent to solve the optimization problem of Eq. (5). The gradient of Eq. (6) with respect to $\boldsymbol{v}_l$ is

$$\begin{aligned} \nabla_{\boldsymbol{v}_l} \mathcal{E} = 2 \sum\nolimits_{n=1}^{N} & p(c_l|\boldsymbol{x}_n^m) \{ p(c_m|\boldsymbol{x}_n^m) - p_{nl} \\ & - \sum\nolimits_{j=1}^{L} [p(c_j|\boldsymbol{x}_n^m) - p_{nj}] p(c_j|\boldsymbol{x}_n^m) \} \boldsymbol{x}_n^m + 2\lambda \boldsymbol{v}_l. \end{aligned} \tag{7}$$

Then $\boldsymbol{v}_l$ $(l = 1, 2, \ldots, L)$ can be iteratively updated by $\boldsymbol{v}_l \leftarrow \boldsymbol{v}_l - \zeta \nabla_{\boldsymbol{v}_l} \mathcal{E}$, where ζ is the step size.

4.4 Baseline D: CNN Regression

Convolutional neural network regression (CNNR) [37] was used for emotion distribution prediction in [11]. We follow the settings of CNNR in [11] as a baseline. That is, a regressor is trained for each emotion category with the exact CNN in [37] except that the number of output nodes is changed to 1 to predict a real value and that the softmax loss layer is replaced with the Euclidean loss layer. The predicted probabilities of all emotion categories are normalized to sum to 1.

4.5 Algorithm E: Shared Sparse Learning

The basic idea of shared sparse learning is that $\boldsymbol{y}^m$ and $\widehat{\boldsymbol{p}}$ can be written in terms of bases $\boldsymbol{X}^m \in \mathbb{R}^{d_m \times N}$ and $\boldsymbol{P} \in \mathbb{R}^{L \times N}$ respectively, but with shared sparse coefficients $\boldsymbol{\theta}^m \in \mathbb{R}^N$. That is

$$\boldsymbol{y}^m = \boldsymbol{X}^m \boldsymbol{\theta}^m \quad \text{and} \quad \widehat{\boldsymbol{p}} = \boldsymbol{P} \boldsymbol{\theta}^m, \tag{8}$$

where $\boldsymbol{\theta}^m$ is obtained by

$$\begin{aligned} \boldsymbol{\theta}^{m*} &= \min \|\boldsymbol{y}^m - \boldsymbol{X}^m \boldsymbol{\theta}^m\|_2^2 + \eta \|\boldsymbol{\theta}^m\|_0, \\ &\text{s.t. } \boldsymbol{\theta}^m \geq \mathbf{0} \text{ and } \|\boldsymbol{\theta}^m\|_1 = 1. \end{aligned} \tag{9}$$

η is a regularization coefficient, similar to λ in Eq. (6). The constraints $\boldsymbol{\theta}^m \geq \mathbf{0}$ and $\|\boldsymbol{\theta}^m\|_1 = 1$ ensure that the predicted $\widehat{\boldsymbol{p}}$ is a probability distribution, that is, $\widehat{\boldsymbol{p}} \geq \mathbf{0}$ and $\|\widehat{\boldsymbol{p}}\|_1 = 1$.

Please note that sparse learning was previously used in many applications [28], [60], [61]. The difference is that the proposed SSL utilizes ℓ_0 norm instead of ℓ_1 norm and is optimized with constraints. Directly optimizing ℓ_1 norm without constraints cannot guarantee that the predicted results sum to 1. Though we may normalize the results to satisfy the probability definition, the emotion correlations are actually ignored.

The optimization of Eq. (9) is a NP-hard problem [74], which cannot be directly solved. By replacing ℓ_0 norm with ℓ_p norm as in [21], [22], we relax the objective function to

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}^m) &= \|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 + \eta\|\boldsymbol{\theta}^m\|_p^p \\ &= \|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 + \eta\sum\nolimits_{n=1}^{N}|\theta_n^m|^p,\end{aligned} \tag{10}$$

where $0 < p \leq 1$. By iteratively reweighted least squares [21], [22], Eq. (10) can be reduced to the following quadratic function with respect to $\boldsymbol{\theta}^m$

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}^m) &= \|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 + \eta\sum_{n=1}^{N}\frac{1}{|\theta_n^m|^{2-p} + \varepsilon}|\theta_n|^2 \\ &= (\boldsymbol{\theta}^m)^{\mathrm{T}}((\boldsymbol{X}^m)^{\mathrm{T}}\boldsymbol{X}^m + \eta\boldsymbol{\Gamma})\boldsymbol{\theta}^m - 2(\boldsymbol{y}^m)^{\mathrm{T}}\boldsymbol{X}^m\boldsymbol{\theta}^m,\end{aligned} \tag{11}$$

where $\varepsilon > 0$ is introduced to avoid division by zero, $\boldsymbol{\Gamma}$ is a diagonal matrix with $\boldsymbol{\Gamma}(n,n) = \frac{1}{|\theta_n^m|^{2-p}+\varepsilon}$. The original optimization problem Eq. (9) is now approximately solved by a series of constrained least squares problems

$$\begin{aligned}&\min\{(\boldsymbol{\theta}^m)^{\mathrm{T}}((\boldsymbol{X}^m)^{\mathrm{T}}\boldsymbol{X}^m + \eta\boldsymbol{\Gamma})\boldsymbol{\theta}^m - 2(\boldsymbol{y}^m)^{\mathrm{T}}\boldsymbol{X}^m\boldsymbol{\theta}^m\}, \\ &\text{s.t. } \boldsymbol{\theta}^m \geq \mathbf{0} \text{ and } \|\boldsymbol{\theta}^m\|_1 = 1.\end{aligned} \tag{12}$$

In practice, $p \rightarrow 0$. The procedure is summarized in Algorithm 1. The computation complexity is $O(c \cdot E \cdot N \cdot d_m)$, where c is the number of iterations in conjugate gradient.

Algorithm 1. Procedure for Shared Sparse Learning

```
Input: Training examples $(X^m, P)$, test feature $\boldsymbol{y}^m$, max-epochs
    $E$, error threshold $\tau$, regularization coefficient $\eta$
Output: Predicted emotion distribution $\widehat{\boldsymbol{p}}$ for $\boldsymbol{y}^m$
Initialization: $\boldsymbol{\theta}^{(0)} = [\theta_1^{(0)}, \ldots, \theta_N^{(0)}] \leftarrow \mathbf{1}/N$, $\varepsilon \leftarrow 10^{-9}$, $p \leftarrow 0$;
    for $e \leftarrow 1$ to $E$ do
    $\boldsymbol{\Gamma}^{(e)}(n,n) \leftarrow \frac{1}{|\theta_n^{(e-1)}|^{2-p}+\varepsilon}, n = 1, \ldots, N$;
    $\boldsymbol{\theta}^{(e)} \leftarrow \min\{\boldsymbol{\theta}^{\mathrm{T}}((\boldsymbol{X}^m)^{\mathrm{T}}\boldsymbol{X}^m + \eta\boldsymbol{\Gamma}^{(e)})\boldsymbol{\theta} - 2(\boldsymbol{y}^m)^{\mathrm{T}}\boldsymbol{X}^m\boldsymbol{\theta}\}$, s.t. $\boldsymbol{\theta} \geq \mathbf{0}, \|\boldsymbol{\theta}\|_1 = 1$
    if $\|\boldsymbol{\theta}^{(e)} - \boldsymbol{\theta}^{(e-1)}\|_2 < \tau$ then
        break;
    end
end
return $\widehat{\boldsymbol{p}} = P\boldsymbol{\theta}^{(e)}$.
```

4.6 Algorithm F: Weighted Multi-Feature Shared Sparse Learning

As shown in [36], image emotions are conveyed by complex visual features from low-level to high-level, such as color contrast and semantic concepts. In practice, we can extract multiple visual features to represent images. Jointly combining the strength of multi-features may improve the prediction performance. CNNR is based on CNN features, while GW, KNNW, SR, and SSL can simply adopt early fusion, late fusion, and canonical correspondence analysis (CCA) fusion [75], [76] to handle multi-features. But they ignore the latent correlation between different features. We present weighted multi-feature shared sparse learning to provide additional useful information to the prediction problem by the constraint of joint sparsity across different features, which may enforce the robustness in coefficient estimation [61].

Algorithm 2. Procedure for Weighted Multi-Feature Shared Sparse Learning

```
Input: Training examples $(\mathbf{X}, P)$, test feature $Y$, max-epochs
    $E$, error threshold $\tau_1, \tau_2$, regularization coefficients
    $\alpha, \beta$
Output: Predicted emotion distribution $\widehat{\boldsymbol{p}}$ for $Y$
Initialization: $\boldsymbol{\theta}^{m(0)} \leftarrow \mathbf{1}/N(m = 1, 2, \ldots, M)$, $\varepsilon \leftarrow 10^{-9}$, $p \leftarrow 0$,
$\boldsymbol{w}^{(0)} \leftarrow \mathbf{1}/M$;
for $e \leftarrow 1$ to $E$ do
    /* Updating $\boldsymbol{\Theta}$ when fixing $\boldsymbol{W}$                /*
    for $m \leftarrow 1$ to $M$ do
        Compute the diagonal matrix $\Phi^{(e)}$ by $\varphi_n^{(e)} \leftarrow 1/$
        $\left(\sqrt{\sum_{m=1}^{M}(\theta_n^{m(e-1)})^2} + \varepsilon\right)$, $\Phi^{(e)}(n,n) \leftarrow \sqrt{\varphi_n^{(e)}}(1 \leq n \leq N)$;
        Optimize $\boldsymbol{\theta}^m$ by

            $\boldsymbol{\theta}^{m(e)} \leftarrow \min w_m^{(e-1)}\|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 + \alpha\|\Phi^{(e)}\boldsymbol{\theta}^m\|_2^2,$
            s.t. $\boldsymbol{\theta}^m \geq \mathbf{0}, \|\boldsymbol{\theta}^m\|_1 = 1;$

    end
    /* Updating $\boldsymbol{w}$ when fixing $\boldsymbol{\Theta}$                /*
    Optimize $\boldsymbol{w}$ by

        $\boldsymbol{w}^{(e)} \leftarrow \min\sum\nolimits_{m=1}^{M} w_m\|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^{m(e)}\|_2^2 + \beta\|\boldsymbol{w}\|_2^2,$
        s.t. $\boldsymbol{w} \geq 0, \|\boldsymbol{w}\|_1 = 1;$

    if $\|\boldsymbol{\theta}^{m(e)} - \boldsymbol{\theta}^{m(e-1)}\|_2 < \tau_1(m = 1, 2, \ldots, M)$ & $\|\boldsymbol{w}^{(e)} - \boldsymbol{w}^{(e-1)}$
    $\|_2 < \tau_2$ then
        break;
    end
end
$\boldsymbol{\Theta}^{(e)} = [\boldsymbol{\theta}^{1(e)}, \boldsymbol{\theta}^{2(e)}, \ldots, \boldsymbol{\theta}^{M(e)}]$;
return $\widehat{\boldsymbol{p}} = \boldsymbol{P}\boldsymbol{\Theta}^{(e)}\boldsymbol{w}^{(e)}$.
```

Extended from SSL, WMFSSL assumes that multi-features Y and $\widehat{\boldsymbol{p}}$ can be written in terms of bases $\mathbf{X}$ and $P \in \mathbb{R}^{L \times N}$ respectively, but with shared sparse coefficients $\boldsymbol{\Theta} \in \mathbb{R}^{N \times M}$. That is

$$\boldsymbol{y}^m = \boldsymbol{X}^m\boldsymbol{\theta}^m(m = 1, 2, \ldots, M) \quad \text{and} \quad \widehat{\boldsymbol{p}} = \boldsymbol{P}\boldsymbol{\Theta}\boldsymbol{w}, \tag{13}$$

where $\boldsymbol{\Theta} = [\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^M]$ and $\boldsymbol{w} = [w_1, w_2, \ldots, w_M]^{\mathrm{T}}$ are obtained by

$$\begin{aligned}[\boldsymbol{\Theta}^*, \boldsymbol{w}^*] = \min\sum_{m=1}^{M} w_m\|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 \\ + \alpha\|\boldsymbol{\Theta}\|_{2,1} + \beta\|\boldsymbol{w}\|_2^2, \\ \text{s.t. } \boldsymbol{\theta}^m \geq \mathbf{0}, \|\boldsymbol{\theta}^m\|_1 = 1 \text{ and } \boldsymbol{w} \geq 0, \|\boldsymbol{w}\|_1 = 1.\end{aligned} \tag{14}$$

α and β are regularization coefficients. The constraints $\boldsymbol{\theta}^m \geq \mathbf{0}$, $\|\boldsymbol{\theta}^m\|_1 = 1$ and $\boldsymbol{w} \geq 0$, $\|\boldsymbol{w}\|_1 = 1$ together ensure that the predicted $\widehat{\boldsymbol{p}}$ is a probability distribution.

Please note that the difference between our method and [61] is that our method adopts a weighted strategy for the integration of different features., and is optimized with constraints to guarantee the property of probability distribution. The method in [61], mainly used for visual classification, cannot be directly employed in our DPD prediction task.

To solve the dual-optimization problem in Eq. (14), we alternatively conduct optimization.

1) Updating $\boldsymbol{\Theta}$ when fixing $\boldsymbol{w}$. Similar to SSL, we employ IRLS [21], [22] to optimize $\boldsymbol{\Theta}$ in Eq. (14), the component $\|\boldsymbol{\Theta}\|_{2,1}$ of which is transformed by

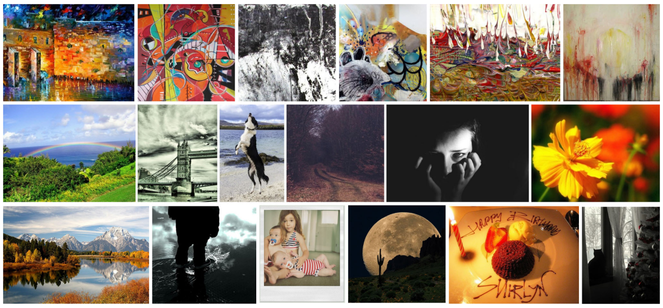


Fig. 4. Image examples that contain DPD information in Abstract (top), Emotion6 (middle) and IESN (bottom) datasets.

$$\|\mathbf{\Theta}\|_{2,1} = \sum_{n=1}^{N}\sqrt{\sum_{m=1}^{M}(\theta_n^m)^2} \simeq \sum_{n=1}^{N}\frac{\sum_{m=1}^{M}(\theta_n^m)^2}{\sqrt{\sum_{m=1}^{M}(\theta_n^m)^2+\varepsilon}}, \qquad (15)$$

where $\varepsilon > 0$ is introduced to avoid division by zero. Let $\varphi_n = 1/(\sqrt{\sum_{m=1}^{M}(\theta_n^m)^2}+\varepsilon)$. Define diagonal matrix $\Phi(n,n) = \sqrt{\varphi_n}(1 \leq n \leq N)$. Then the objective function of Eq. (14) with respect to $\mathbf{\Theta}$ is transformed to

$$\mathcal{O}(\mathbf{\Theta}) = \sum_{m=1}^{M} w_m\|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 + \alpha\|\Phi\boldsymbol{\theta}^m\|_2^2. \qquad (16)$$

$\min\mathcal{O}(\mathbf{\Theta})$ can be optimized for each $\boldsymbol{\theta}^m$ independently

$$\begin{aligned} &\min\, w_m\|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 + \alpha\|\Phi\boldsymbol{\theta}^m\|_2^2, \\ &\text{s.t. } \boldsymbol{\theta}^m \geq \mathbf{0}, \|\boldsymbol{\theta}^m\|_1 = 1, \end{aligned} \qquad (17)$$

which can be easily and efficiently solved by off-the-shelf quadratic optimization methods.

2) Updating $\boldsymbol{w}$ *when fixing* $\mathbf{\Theta}$. The optimization problem of Eq. (14) with respect to $\boldsymbol{w}$ is transformed to

$$\begin{aligned} &\min\sum_{m=1}^{M} w_m\|\boldsymbol{y}^m - \boldsymbol{X}^m\boldsymbol{\theta}^m\|_2^2 + \beta\|\boldsymbol{w}\|_2^2, \\ &\text{s.t.}\ \ \boldsymbol{w} \geq 0, \|\boldsymbol{w}\|_1 = 1, \end{aligned} \qquad (18)$$

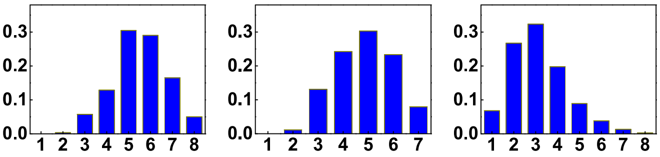
which is a also quadratic programming problem. The learning procedure is summarized in Algorithm 2. The computation complexity is $O(c \cdot E \cdot N \cdot \sum_{m=1}^{M} d_m)$, where c is the number of iterations in conjugate gradient when optimizing $\boldsymbol{\theta}^m$. As c, E, N and d_m are all not large in experiment, the algorithm can be converged in a few seconds.

5 Experiment Setup

To the best of our knowledge, there are three public datasets that contain DPD information of image emotions: Abstract [5], Emotion6 [11] and IESN [13], [18]. In this section, we introduce these benchmark datasets and the experimental settings in evaluating the performance of emotion distribution prediction.

5.1 Datasets

The Abstract dataset [5] includes 279 abstract paintings without any recognizable objects. These images were peer rated in a web-survey by approximately 230 people into 8 emotion categories, including *anger*, *disgust*, *fear*, *sadness* as negative emotions and *amusement*, *awe*, *contentment*, *excitement* as positive emotions, based on a rigorous psychological study [27]. On average each image was rated about 14 times [5]. Only 228 images can be used for affective image classification [5], [10], while all the 279 images can be used for emotion distribution prediction.



(a) Abstract (b) Emotion6 (c) IESN

Fig. 5. The distribution of images that are labeled with different emotion numbers, where the horizontal axis is the number of different emotions, and the vertical axis is image proportion. The majority of images are labeled with at least two emotion categories, which demonstrates that the perceived emotions are truly subjective.

The Emotion6 dataset [11] consists of 1,980 images collected from Flickr, 330 for each of the Ekman's 6 basic emotions [26]: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. The emotional responses from subjects were obtained using Amazon Mechanical Turk (AMT). Each image was scored by 15 subjects into Ekman's 6 basic emotions and *neutral*.

The Image-Emotion-Social-Net dataset [13], [18], which contains 1,012,901 images collected from Flickr keywords based searching strategy [8], [9], [30], is firstly used for personalized emotion prediction. Employing viewers to label the emotions of images is tedious and time-consuming, and even impossible for large-scale datasets. Instead, the emotion information of the social images in IESN is automatically obtained from the text data, such as the metadata and comments, based on the assumption that viewers express their emotion perceptions of the images by the text comments. Similar to Abstract [5], the emotions are also classified into 8 categories. Totally, we select 3,792 images, each of which is assigned with more than 15 categorial emotion labels.

Some image examples in the three datasets are illustrated in Fig. 4. The emotion distribution on the 8 or 7 categories of each image can be easily obtained by normalization, i.e., dividing the number of subjects who perceive each emotion category by the number of all emotion perceptions. For example, given an image, suppose the perceived emotion number by 20 subjects on the 8 emotion categories is $v = [7, 0, 4, 5, 0, 6, 2, 1]$, then the DPD is $v/\sum(v) = [0.28, 0, 0.16, 0.2, 0, 0.24, 0.08, 0.04]$. Note that one subject can perceive multiple emotions from the same image. The distribution of emotion numbers for the images in the three datasets is shown in Fig. 5, from which we can clearly see the subjectivity issue of emotion perceptions. Please note that the emotions in IESN is less subjective than those in Abstract and Emotion6. This is probably because that the images in Abstract are abstract paintings without clear semantics, the images with apparent expressions or text directly related to emotions are removed when constructing Emotion6, while the images in IESN are social ones and the emotions of one viewer may be easily influenced by another.

5.2 Emotion Features

The features that determine the emotions of an image may vary for different kinds of images [36]. To enhance the representation power of visual features, we extract various features, including hand-crafted ones of different levels and learning-based ones.

We first extract two classes of low-level hand-crafted features for their global descriptors of the overall image content. The first is the GIST feature, which is one of the most commonly used features, for its relatively powerful description ability of visual phenomena in a scene perspective [77]. The second class includes the features derived from elements-of-art, i.e. color and texture [5].

Mid-level features are more semantic, interpretable and have stronger link to emotions than low-level features [10]. Here we extract two classes of mid-level features. The first is attribute based representation, including 102 dimensional attributes which are commonly used by humans to describe scenes [77]. Features inspired from principles-of-art, including *balance*, *contrast*, *harmony*, *variety*, *gradation*, and *movement* [10], are extracted as another mid-level features.

High-level features are the detailed semantic contents contained in images. People can easily understand the emotions conveyed in images by recognizing the semantics. We extract a set of concepts described by the 1,200 adjective noun pairs (ANPs) [9], which are detected by a large detector library SentiBank [9].

Further, we extract the deep features from the response of the fully connected layer (FC) 7 of the AlexNet trained on ImageNet [37], which is the final fully connected layer before producing the class predictions. The deep feature for each image is a 4096-dimensional vector.

The six sets of extracted visual features are abbreviated as GIST, Elem, Attr, Prin, ANP and CNN with dimension 512, 48, 102, 165, 1200 and 4096, respectively.

5.3 Evaluation Metrics

The sum of squared difference, the Kullback-Leibler divergence, the Bhattacharyya coefficient and the coefficient of determination are used as the evaluation metrics. Suppose $\mathbf{T}$ is the test set. For test image I, the ground-truth emotion distribution is $\boldsymbol{p} = [p_1, p_2, \ldots, p_L]^{\mathrm{T}}$ and the predicted emotion distribution is $\widehat{\boldsymbol{p}} = [\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_L]^{\mathrm{T}}$.

The sum of squared difference (SSD) of test image I is defined as

$$SSD_{\mathrm{I}}(I) = \sum_{l=1}^{L} (\widehat{p}_l - p_l)^2;$$

the SSD of emotion c_l is defined as

$$SSD_{\mathrm{E}}(c_l) = \frac{1}{\#\mathbf{T}} \sum\nolimits_{I\in\mathbf{T}} (\widehat{p}_l - p_l)^2,$$

where $\#\mathbf{T}$ is the number of test images; the overall SSD on $\mathbf{T}$ is defined as

$$SSD = \frac{1}{\#\mathbf{T}} \sum_{I\in\mathbf{T}} SSD_{\mathrm{I}}(I) \quad \text{or} \quad \sum_{l=1}^{L} SSD_{\mathrm{E}}(c_l).$$

SSD ranges from 0 to 1 and a good prediction results in a small SSD value.

As a classical measure of distance between distributions, the KL divergence (KL)[2] of the predicted distribution $\widehat{\boldsymbol{p}}$ from the ground truth distribution $\boldsymbol{p}$ is defined as

$$KL(\boldsymbol{p}||\widehat{\boldsymbol{p}}) = \sum_{l=1}^{L} \left(p_l \ln p_l - p_l \ln \widehat{p}_l\right).$$

KL measures the amount of information lost when $\widehat{\boldsymbol{p}}$ is used to approximate $\boldsymbol{p}$. $KL \geq 0$ and lower value indicates better performance, with equality if, and only if the predicted distribution $\widehat{\boldsymbol{p}}$ is equal to the ground truth distribution $\boldsymbol{p}$. Since KL is not well defined when a bin has value 0, we use a small value 10^{-10} to approximate the values in such bins [11].

The Bhattacharyya coefficient (BC)[3] between two DPDs $\boldsymbol{p}$ and $\widehat{\boldsymbol{p}}$ is defined by

$$BC(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \sum\nolimits_{l=1}^{L} \sqrt{p_l \widehat{p}_l}.$$

$0 \leq BC \leq 1$ and larger value represents better results.

The coefficient of determination, denoted R^2 statistic (R^2),[4] between two DPDs $\boldsymbol{p}$ and $\widehat{\boldsymbol{p}}$ is defined by

$$R^2(\boldsymbol{p}, \widehat{\boldsymbol{p}}) = \frac{\mathrm{cov}^2(\boldsymbol{p}, \widehat{\boldsymbol{p}})}{\mathrm{var}(\boldsymbol{p})\mathrm{var}(\widehat{\boldsymbol{p}})},$$

where $\mathrm{cov}(.,.)$ and $\mathrm{var}(.)$ are the covariance function and variance function, respectively. R^2 ranges from 0 to 1. If $\widehat{\boldsymbol{p}}$ perfectly matches $\boldsymbol{p}$, the R^2 value is 1. If there is no linear relationship between the two DPDs, R^2 is 0.

Please note that (1) SSD measures the performance from the aspect of regression, while KL, BC and R^2 measure the distance between two distributions; (2) KL and BC emphasize on each individual element, whereas R^2 considers the variance among all the elements in the DPD. Note that other similarity measures between distributions, such as the earth mover's distance and Chebyshev distance [11], can also be used as evaluation metrics. Due to the page limit, we do not report these results.

5.4 Implementation Details

We randomly select 80, 50 and 50 percent images from the Abstract, Emotion6 and IESN datasets respectively as the training set and the remained form the testing set. For KNNW, K is empirically set to 200, 500 and 500 respectively for the Abstract, Emotion6 and IESN datasets. For SR, λ is decided by 5-fold cross validation in the training set. For CNNR, as in [11], the AlexNet [37] is firstly pre-trained using the Caffe reference model [78] and then fine-tuned with our training set. As in [11], the number of output nodes is changed to 1 to predict a real value. The softmax loss layer is replaced with the Euclidean loss layer. In the predicting phase, the probabilities of all emotion categories are normalized to sum to 1. The following parameter settings are adopted: $\eta = 0.0001$ for SSL, $\alpha = 0.05$ and $\beta = 0.1$ for WMFSSL. We also conduct empirical analysis on parameter sensitivity, which demonstrates that SSL and WMFSSL have superior and stable performances with a wide range of parameter values on all three datasets. The features that are over 50-dimensional before fusion are reduced to 50 by PCA to accelerate the optimization. The settings of early and late fusions are similar to [72]. For early fusion, we concatenate the different features after normalization of each feature and then put the combined one into the learning algorithms. For late fusion, we first predict

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

[3] https://en.wikipedia.org/wiki/Bhattacharyya_distance

[4] https://en.wikipedia.org/wiki/Coefficient_of_determination

TABLE 1
Performance Comparison on Abstract Dataset Measured by SSD, KL, BC, R^2 ($\times10^{-1}$) and the Standard Deviations ($\times10^{-1}$)

		GIST	Elem	Attr	Prin	ANP	CNN	Early	Late	CCA
	GW	1.382 (0.060)	1.365 (0.056)	**1.387** (0.060)	*1.352* (0.039)	1.370 (0.056)	1.371 (0.056)	1.369 (0.056)	*1.368* (0.055)	1.384 (0.057)
S	KNNW	1.505 (0.113)	1.397 (0.979)	1.507 (0.096)	1.361 (0.040)	1.389 (0.051)	*1.257* (0.050)	1.263 (0.049)	*1.250* (0.047)	1.350 (0.061)
S	SR	1.432 (0.068)	2.959 (1.410)	8.451 (1.911)	2.169 (0.299)	3.904 (1.817)	*1.405* (0.616)	1.426 (0.083)	1.970 (0.490)	*1.405* (0.044)
D	CNNR	–	–	–	–	–	**1.244** (0.096)	–	–	–
	SSL	**1.369** (0.073)	**1.346** (0.134)	1.473 (0.090)	**1.316** (0.106)	**1.354** (0.035)	*1.282* (0.046)	1.271 (0.045)	1.241 (0.055)	*1.227* (0.060)
	WMFSSL	–	–	–	–	–	–		***1.191*** (0.060)	
	GW	5.543 (0.168)	5.472 (0.160)	**5.558** (0.163)	*5.430* (0.128)	5.497 (0.150)	5.495 (0.157)	5.491 (0.155)	*5.490* (0.152)	5.549 (0.161)
	KNNW	6.113 (0.379)	5.687 (0.362)	6.166 (0.370)	5.576 (0.247)	5.657 (0.362)	*5.173* (0.176)	5.142 (0.132)	*5.105* (0.156)	5.491 (0.146)
K	SR	5.754 (0.183)	36.61 (39.93)	98.01 (112.4)	15.35 (2.329)	28.86 (15.38)	*5.653* (0.160)	5.742 (0.235)	7.074 (1.069)	*5.613* (0.858)
L	CNNR	–	–	–	–	–	**5.103** (0.305)	–	–	–
	SSL	**5.525** (0.329)	**5.439** (1.379)	6.070 (0.204)	**5.421** (0.551)	**5.475** (0.153)	*5.225* (0.177)	5.126 (0.096)	5.034 (0.177)	*5.014* (0.141)
	WMFSSL	–	–	–	–	–	–		***4.820*** (0.209)	
	GW	7.984 (0.072)	8.010 (0.072)	**7.980** (0.071)	*8.017* (0.072)	7.996 (0.070)	7.998 (0.069)	7.999 (0.069)	*7.999* (0.070)	7.982 (0.070)
	KNNW	7.898 (0.125)	8.050 (0.078)	7.880 (0.102)	8.103 (0.069)	8.039 (0.099)	*8.111* (0.077)	8.120 (0.048)	*8.150* (0.068)	8.081 (0.082)
B	SR	7.898 (0.069)	6.724 (1.218)	4.332 (1.247)	6.982 (0.236)	6.355 (0.818)	*7.928* (0.061)	7.913 (0.073)	7.683 (0.221)	*7.956* (0.247)
C	CNNR	–	–	–	–	–	**8.173** (0.093)	–	–	–
	SSL	**7.992** (0.108)	**8.106** (0.168)	7.849 (0.928)	**8.118** (0.077)	**8.095** (0.048)	*8.118* (0.074)	8.142 (0.046)	8.210 (0.081)	*8.268* (0.039)
	WMFSSL	–	–	–	–	–	–		***8.319*** (0.078)	
	GW	1.863 (0.062)	2.027 (0.120)	1.817 (0.065)	*2.066* (0.092)	1.865 (0.094)	1.945 (0.109)	*1.931* (0.102)	*1.931* (0.081)	1.897 (0.092)
	KNNW	**1.919** (0.365)	**2.316** (0.557)	1.758 (0.403)	2.447 (0.270)	2.327 (0.556)	*2.612* (0.227)	2.759 (0.341)	*2.789* (0.328)	2.416 (0.375)
R^2	SR	1.754 (0.200)	1.400 (0.519)	**1.891** (0.697)	1.790 (0.343)	1.674 (0.506)	*2.122* (0.232)	1.632 (0.627)	1.534 (0.547)	*1.687* (0.293)
	CNNR	–	–	–	–	–	**2.796** (0.309)	–	–	–
	SSL	1.915 (0.236)	2.161 (0.372)	1.850 (0.395)	**2.478** (0.428)	**2.483** (0.337)	*2.660* (0.385)	2.678 (0.354)	2.818 (0.456)	*2.930* (0.145)
	WMFSSL	–	–	–	–	–	–		***2.993*** (0.467)	

the emotion distribution using each feature and then compute the fused distribution using feature distance-based linear weighting. The settings of CCA fusion are similar to [76]. The linear combinations are transformed by maximizing the pairwise correlations across two features. Feature-level fusion is performed by concatenation of the transformed feature vectors for every two features successively. CCA fusion adopted here is, in fact, another kind of early fusion. For better comparison, the parameters of the baselines are carefully tuned and the best results are reported. To remove the influence of any randomness, we perform 20 runs and report the average results with the standard deviation.

6 Experimental Results

In this section, we report the results on different uni-features, feature fusion methods, and parameter sensitivity.

6.1 On Uni-Features

Firstly, we conduct experiments to compare the performance of different visual features and uni-feature based methods for emotion distribution prediction. The average performances measured by SSD, KL, BC, R^2 and the standard deviations on Abstract, Emotion6 and IESN datasets are summarized in Tables 1, 2, and 3, respectively. The best uni-feature for each learning method is shown in italic, while the best method for each uni-feature is emphasized in bold.

From the results, we have the following observations. (1) Generally, the CNN features have stronger, at least comparable, discriminability than the hand-crafted ones; the high-level and mid-level hand-crafted features perform better than low-level ones. These results are consistent with several existing literatures [13], [14], [36]. (2) The CNNR method achieves the best results in most cases with uni-features, which demonstrates the effectiveness of CNNR in DPD prediction of image emotions [11]. (3) For hand-crafted visual features, the proposed SSL method outperforms the baselines, including GW, KNNW and SR. (4) The performance of SR is not stable. Though direct and simple, GW and KNNW perform better than SR. (5) The metrics SSD, KL and BC are more consistent to measure the performance of distribution prediction than R^2.

Besides the common observations above, there are still some different results across datasets. (1) The features derived from principles-of-art and elements-of-art perform even better than the high-level ANP features on Abstract and Emotion6 datasets. This is probably because the images in Abstarct are abstract paintings without recognizable objects, the emotions of which are mainly evoked by art theory and aesthetics. Meanwhile, the apparent semantics directly related to the evoked emotions, such as expressive faces, are removed in the Emotion6 dataset construction [11]. (2) For most uni-features, GW performs better than KNNW on Abstract dataset while reversely on Emotion6 and IESN datasets, which might be caused by the fact that the number of bases in Abstract dataset is small, which cannot well represent the variability within each emotion. (3) The metric R^2 is much larger in IESN dataset than Abstract and Emotion6 datasets, since the evoked emotion numbers of each image is relatively smaller in IESN (Fig. 5) due to the influence of social factors, such as the joined interest groups.

6.2 On Different Feature Fusion Methods

Secondly, we compare the performance of different feature fusion methods for emotion distribution prediction, including the proposed WMFSSL, early fusion, late fusion and CCA fusion for GW, KNNW, SR, and SSL. On the right of

TABLE 2
Performance Comparison on Emotion6 Dataset Measured by SSD, KL, BC, R^2 ($\times 10^{-1}$) and the Standard Deviations ($\times 10^{-1}$)

		GIST	Elem	Attr	Prin	ANP	CNN	Early	Late	CCA
	GW	**1.991** (0.043)	1.959 (0.052)	1.992 (0.043)	*1.914* (0.050)	1.972 (0.046)	1.966 (0.047)	1.964 (0.047)	1.963 (0.047)	*1.960* (0.047)
S	KNNW	2.123 (0.048)	1.871 (0.084)	2.161 (0.052)	1.844 (0.057)	1.850 (0.014)	*1.448* (0.075)	1.540 (0.017)	1.616 (0.008)	*1.533* (0.036)
S	SR	*2.226* (0.063)	5.336 (4.246)	2.319 (0.172)	5.220 (4.156)	5.878 (2.473)	2.787 (0.245)	4.690 (1.383)	*2.212* (0.121)	3.562 (0.185)
D	CNNR	–	–	–	–	–	**1.394** (0.080)	–	–	–
	SSL	2.043 (0.061)	**1.828** (0.061)	**1.984** (0.033)	**1.806** (0.065)	**1.794** (0.122)	*1.427* (0.043)	1.344 (0.033)	1.402 (0.086)	*1.332* (0.042)
	WMFSSL	–	–	–	–	–	–		***1.268*** (0.076)	
	GW	**6.183** (0.016)	6.098 (0.044)	**6.189** (0.017)	*5.965* (0.033)	6.121 (0.027)	6.109 (0.026)	6.099 (0.026)	6.103 (0.027)	*6.067* (0.024)
	KNNW	6.768 (0.083)	6.039 (0.204)	6.871 (0.171)	5.892 (0.291)	5.972 (0.232)	*5.349* (0.222)	5.058 (0.063)	5.067 (0.025)	*5.024* (0.028)
K	SR	*7.093* (0.451)	44.62 (52.25)	7.499 (0.612)	28.31 (29.84)	36.02 (31.69)	11.89 (5.250)	19.93 (0.896)	*6.880* (0.075)	8.379 (0.631)
L	CNNR	–	–	–	–	–	**4.846** (0.469)	–	–	–
	SSL	6.389 (0.317)	**5.999** (0.882)	6.205 (0.096)	**5.863** (0.445)	**5.705** (0.306)	*5.244* (0.041)	4.825 (0.084)	5.064 (0.150)	*4.796* (0.082)
	WMFSSL	–	–	–	–	–	–		***4.793*** (0.097)	
	GW	**7.876** (0.023)	7.892 (0.031)	7.868 (0.023)	*7.935* (0.029)	7.884 (0.025)	7.888 (0.026)	7.890 (0.025)	7.890 (0.026)	*7.912* (0.024)
	KNNW	7.822 (0.031)	**8.056** (0.082)	7.795 (0.044)	8.073 (0.043)	8.118 (0.009)	*8.294* (0.055)	8.365 (0.005)	8.222 (0.008)	*8.374* (0.006)
B	SR	*7.723* (0.033)	5.991 (2.170)	7.524 (0.082)	6.111 (2.004)	5.870 (1.399)	7.270 (0.451)	6.302 (0.167)	*7.744* (0.056)	6.615 (0.154)
C	CNNR	–	–	–	–	–	**8.437** (0.050)	–	–	–
	SSL	7.868 (0.097)	7.940 (0.049)	**7.909** (0.009)	**8.111** (0.113)	**8.151** (0.061)	*8.402* (0.015)	8.484 (0.012)	8.411 (0.044)	*8.502* (0.014)
	WMFSSL	–	–	–	–	–	–		***8.529*** (0.059)	
	GW	2.683 (0.345)	2.792 (0.316)	2.745 (0.367)	*3.166* (0.365)	2.691 (0.322)	2.685 (0.352)	2.691 (0.337)	*2.801* (0.345)	2.736 (0.336)
	KNNW	2.738 (0.331)	3.582 (0.053)	2.505 (0.068)	3.625 (0.015)	3.632 (0.288)	*4.142* (0.021)	4.386 (0.295)	4.012 (0.236)	*4.427* (0.283)
R^2	SR	2.105 (0.138)	2.314 (1.212)	1.682 (0.075)	*2.399* (0.999)	2.256 (1.369)	1.700 (0.075)	2.524 (0.867)	*3.175* (0.232)	2.739 (0.835)
	CNNR	–	–	–	–	–	**4.434** (0.348)	–	–	–
	SSL	**2.755** (0.381)	**3.601** (0.104)	**2.832** (0.151)	**3.644** (0.117)	**3.683** (0.182)	*4.237* (0.014)	4.533 (0.180)	4.368 (0.161)	*4.582* (0.174)
	WMFSSL	–	–	–	–	–	–		***4.679*** (0.172)	

Tables 1, 2 and 3, the better fusion method for GW, KNNW, SR, and SSL is shown in italic, while the best overall result is highlighted in both italic and bold. Comparing the results, we can observe that: (1) fusing multi-features by either early fusion, late fusion or CCA fusion for GW, KNNW, SR, and SSL can obtain better prediction performance than most of uni-features; (2) the best fusion method is dependent on the methods and datasets; on Abstract dataset, late fusion achieves better performance for GW and KNNW, while CCA fusion performs better for SR and SSL; on Emotion6

TABLE 3
Performance Comparison on IESN Dataset Measured by SSD, KL, BC, R^2 ($\times 10^{-1}$) and the Standard Deviations ($\times 10^{-1}$)

		GIST	Elem	Attr	Prin	ANP	CNN	Early	Late	CCA
	GW	2.027 (0.035)	1.899 (0.027)	2.017 (0.035)	1.966 (0.036)	*1.867* (0.036)	1.882 (0.038)	1.872 (0.038)	1.890 (0.035)	*1.853* (0.036)
S	KNNW	2.019 (0.199)	1.900 (0.029)	1.895 (0.003)	1.893 (0.024)	1.786 (0.083)	*1.761* (0.123)	1.706 (0.062)	1.713 (0.051)	*1.688* (0,054)
S	SR	*3.485* (0.440)	5.266 (0.348)	5.247 (0.335)	5.240 (0.332)	5.153 (0.316)	4.781 (1.120)	3.692 (0.349)	3.737 (0.067)	*3.524* (0.328)
D	CNNR	–	–	–	–	–	**1.703** (0.022)	–	–	–
	SSL	**1.928** (0.431)	**1.854** (0.078)	**1.863** (0.008)	**1.852** (0.113)	**1.728** (0.002)	*1.719* (0.054)	1.676 (0.125)	1.706 (0.090)	*1.628* (0.119)
	WMFSSL	–	–	–	–	–	–		***1.569*** (0.014)	
	GW	5.967 (0.095)	5.479 (0.093)	5.356 (0.098)	5.227 (0.110)	*4.947* (0.099)	4.960 (0.101)	4.903 (0.100)	4.932 (0.099)	*4.885* (0.096)
	KNNW	5.824 (0.919)	5.284 (0.115)	5.236 (0.092)	5.195 (0.716)	5.059 (0.139)	*4.916* (0.119)	4.853 (0.566)	4.901 (0.198)	*4.817* (0.542)
K	SR	18.86 (7.174)	14.53 (0.495)	14.57 (0.500)	14.46 (0.486)	*14.35* (0.479)	15.02 (3.598)	14.51 (0.496)	*10.09* (0.078)	12.48 (0.462)
L	CNNR	–	–	–	–	–	**4.828** (0.953)	–	–	–
	SSL	**5.606** (1.136)	**5.292** (0.385)	**5.173** (0.177)	**5.083** (0.929)	**4.915** (0.288)	*4.874* (0.115)	4.812 (0.108)	4.837 (0.134)	*4.783* (0.100)
	WMFSSL	–	–	–	–	–	–		***4.777*** (0.016)	
	GW	7.910 (0.013)	8.258 (0.006)	8.319 (0.013)	8.385 (0.018)	8.457 (0.017)	*8.453* (0.016)	8.455 (0.017)	8.453 (0.014)	*8.461* (0.016)
	KNNW	8.383 (0.086)	8.448 (0.009)	8.449 (0.058)	8.457 (0.006)	8.425 (0.092)	*8.511* (0.096)	8.511 (0.038)	8.508 (0.028)	*8.557* (0.034)
B	SR	*5.947* (0.307)	5.311 (0.089)	5.321 (0.090)	5.342 (0.089)	5.736 (0.087)	5.496 (0.873)	6.125 (0.090)	*6.406* (0.049)	6.392 (0.088)
C	CNNR	–	–	–	–	–	**8.534** (0.047)	–	–	–
	SSL	**8.450** (0.290)	**8.456** (0.054)	**8.461** (0.020)	**8.486** (0.055)	**8.505** (0.003)	*8.515* (0.037)	8.542 (0.072)	8.525 (0.068)	*8.561* (0.066)
	WMFSSL	–	–	–	–	–	–		***8.583*** (0.015)	
	GW	**7.051** (0.284)	6.903 (0.292)	6.998 (0.281)	7.033 (0.273)	7.127 (0.285)	*7.146* (0.281)	7.162 (0.282)	7.156 (0.283)	*7.186* (0.280)
	KNNW	7.014 (0.435)	**7.150** (0.356)	7.119 (0.295)	7.155 (0.213)	7.219 (0.256)	*7.230* (0.063)	7.281 (0.180)	7.252 (0.325)	*7.296* (0.161)
R^2	SR	*6.922* (0.383)	6.324 (0.949)	6.576 (0.321)	6.730 (0.291)	6.831 (0.310)	6.795 (0.641)	6.963 (0.688)	*7.016* (0.407)	6.968 (0.672)
	CNNR	–	–	–	–	–	**7.306** (0.015)	–	–	–
	SSL	6.828 (0.361)	7.043 (0.059)	**7.154** (0.283)	**7.201** (0.501)	**7.221** (0.319)	7.232 (0.239)	*7.314* (0.178)	7.265 (0.171)	*7.335* (0.175)
	WMFSSL	–	–	–	–	–	–		***7.358*** (0.382)	

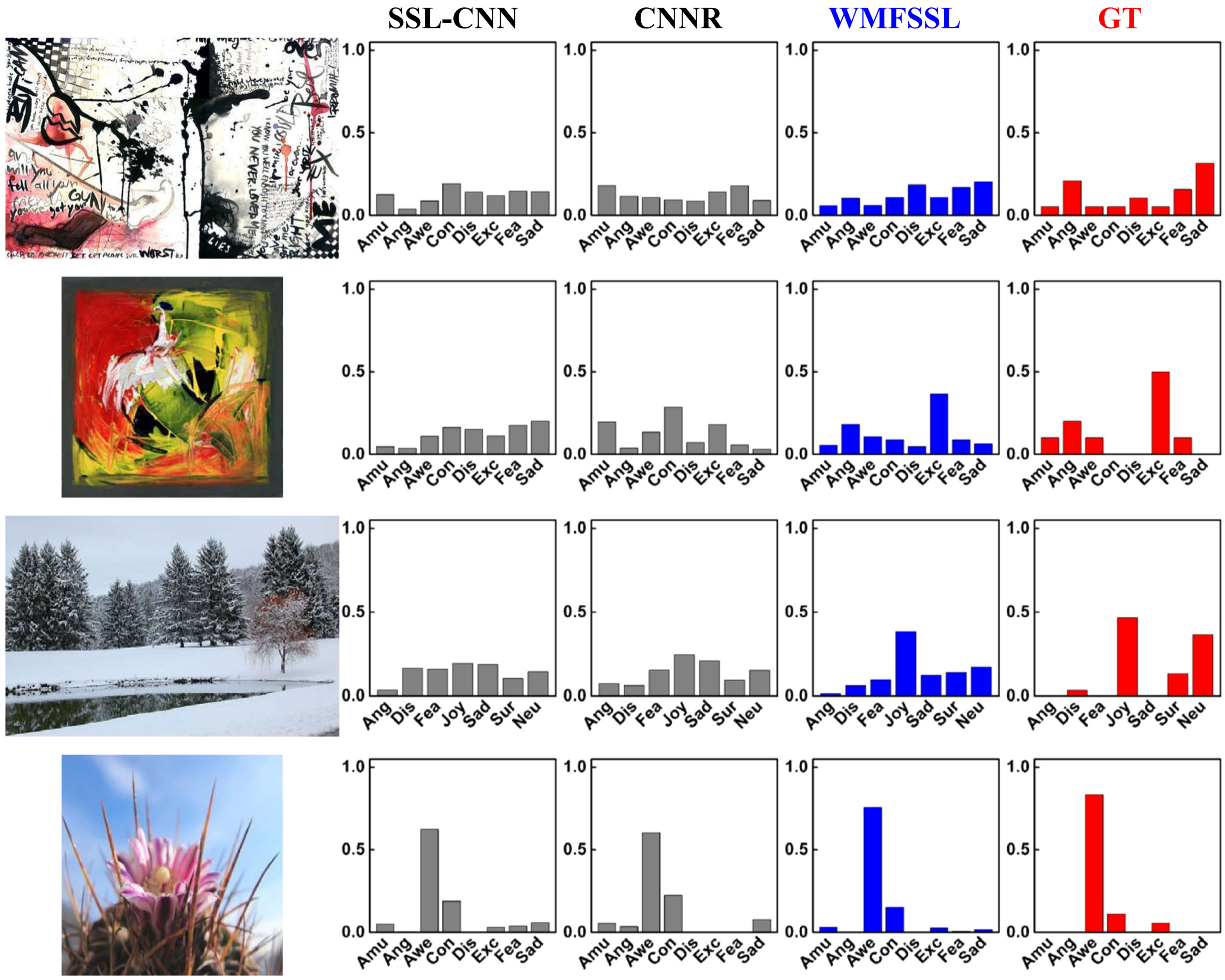


Fig. 6. Predicted emotion distributions using the proposed (WMF)SSL and the best state-of-the-art approach (CNNR [11]). Images and the corresponding ground truth distributions ('GT') are shown in the first and last columns of each group, respectively.

and IESN datasets, CCA fusion works better for GW, KNNW and SSL, while late fusion outperforms early fusion and CCA fusion for SR; (3) SSL with CCA fusion method outperforms CNNR on Abstract, Emotion6 and IESN datasets; (4) generally, CCA fusion works better than early fusion; (5) the proposed fusion method, namely WMFSSL, performs the best on the three datasets, which demonstrates the effectiveness of WMFSSL for emotion distribution prediction.

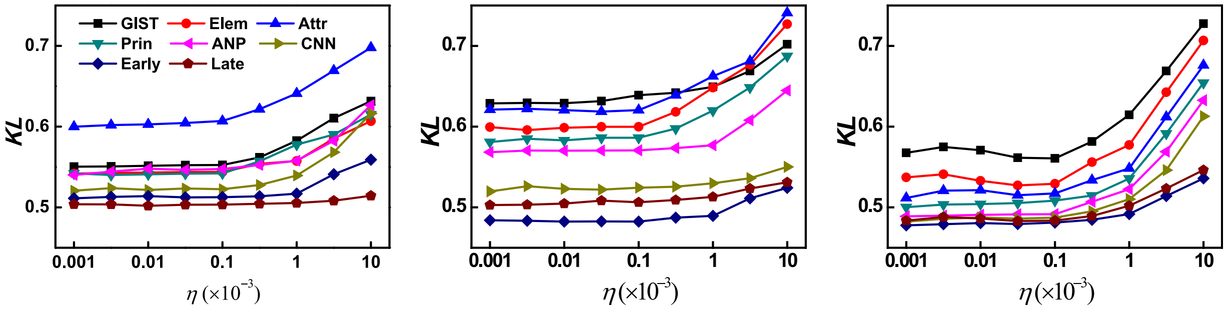
Specifically, the performance improvements of SSL with the best fusion method over the best uni-features measured by SSD, KL, BC, R^2 are 3.20, 3.66, 1.13, 5.94 percent on Abstract, 5.82, 7.99, 0.98, 6.99 percent on Emotion6 and 2.50, 1.48, 0.32, 1.13 percent on IESN datasets, respectively. Compared with the best results of GW, KNNW, SR, CNNR and SSL, WMFSSL achieves the KL performance gains of 11.23, 5.58, 14.74, 5.55, 4.25 percent on Abstract, 19.65, 5.24, 30.33, 1.09, 0.66 percent on Emotion6 and 2.60, 1.57, 52.66, 1.06, 0.73 percent on IESN datasets, respectively. These results demonstrate that the proposed (WMF)SSL method can achieve superior performance over the state-of-the-art approaches for emotion distribution prediction. The performance improvements benefit from the representation complementation of various features jointly explored in the proposed method. Further, the weights of different features are automatically learned, which indicates that the proposed method can easily generalize to new datasets.

Fig. 6 shows the predicted emotion distributions on images from different datasets. For convenience, we just compare the proposed method with the best baseline method, i.e. CNNR [11]. From the results, we can see that WMFSSL generates the most similar distributions to the ground truth, which demonstrates the effectiveness of the proposed method.

We also compare the computational efficiency between the proposed method and the baselines. For fair comparison, we just compute the average test time of each algorithm (GW, KNNW, SR, SSL all with CCA fusion, CNNR, and WMFSSL), excluding the time of feature extraction and dimension reduction. The average computational time of different methods in the test stage is listed in Table 4. The proposed method costs more time than other methods due to its complexity of optimization. Since the dataset size is not large, all the methods can predict a test image's DPD in a few seconds. The above computational time is obtained by conducting the experiments on a server with an Intel

TABLE 4
The Average Computational Time (Seconds) in the Test Stage

	GW	KNNW	SR	CNNR	SSL	WMFSSL
Abstract	2.35e-5	4.23e-5	1.21e-4	2.46e-2	0.051	0.549
Emotion6	6.37e-5	9.89e-5	3.56e-4	3.52e-2	0.168	1.236
IESN	1.19e-4	2.71e-4	4.57e-3	7.68e-2	0.506	5.483



(a) Abstract (b) Emotion6 (c) IESN

Fig. 7. The influence of parameter η in SSL on the three datasets measured by KL divergence.

Core i7-4770K CPU, 32GB RAM and 64-bit ubuntu 16.04 LTS operating system.

6.3 On Parameter Sensitivity

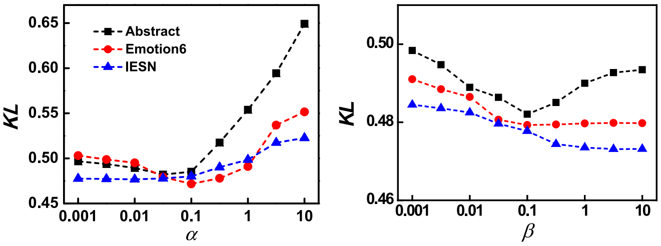
In SSL, we have one sparsity parameter η. In WMFSSL, we have two model parameters, α to control the model sparsity and β as the feature weight parameter. We investigate how sensitive SSL and WMFSSL are to the parameters. When analyzing α and β in WMFSSL, we fix the other as the value we introduced above.

The influence of parameter η in SSL on Abstract, Emotion6 and IESN datasets measured by KL divergence is shown in Fig. 7. Generally, with the decrease of η, the performance becomes better. When η decreases to 0.0001, the performance turns to be stable. The parameter η controls the sparsity of the model. When η is too large, the model tends to be too sparse, which cannot guarantee that the test features are well linearly represented by the training features. In such cases, the prediction performance might be degraded.

The influences of the regularization parameters α, β in WMFSSL are validated, with results shown in Fig. 8. From these results, we can observe that: (1) the influences of α, β are different on different datasets; more stable performances are obtained on Emotion6 and IESN datasets than Abstract dataset; (2) generally, with the decrease of α, the performance tends to become better with relatively stable performance achieved when α decreases to 0.1; (3) on Abstract dataset, with the increase of β, the performance firstly becomes better and then turns worse, meaning that there exists the best β; though not so obviously, WMFSSL achieves better KL values when $\beta \geq 0.1$ on Emotion6 and IESN datasets.

6.4 On Feature Dimension Reduction in WMFSSL

In this subsection, we evaluate the influence of feature dimension reduction by PCA in WMFSSL on the prediction results. The performance comparison with and without PCA is shown in Table 5. From these results, it is clear that after feature dimension reduction by PCA, all the metrics degrade on all the three datasets without a significant performance drop. This is because although the principal components are preserved by PCA, some discriminative information may be missing. On the other hand, we are able to accelerate the computation by about 20X after PCA, which is calculated based on the computation complexity in Section 4.6.



(a) (b)

Fig. 8. The influence of different parameters in WMFSSL on the three datasets measured by KL divergence: (a) the influence of α when $\beta = 0.1$, and (b) the influence of β when $\alpha = 0.05$.

6.5 On Feature Contribution in WMFSSL

Finally, we evaluate the impact of different features in the proposed WMFSSL by comparing the performance of removing one feature each time and using all features. The results are shown in Table 6. It should be noted that the smaller the performance is as compared to WMFSSL, the larger gain the feature contributes to WMFSSL, taking KL for example. From the results, we can conclude that: (1) similar contribution order for each dataset can be obtained as the discriminability order in Tables 1, 2 and 3; CNN contributes more than other features; among the hand-crafted features, Prin and Elem contribute more in Abstract dataset, while ANP and Prin are more discriminative in Emotion6 and IESN datasets, this is probably because the images in Abstract dataset are abstract paintings without obvious semantics, while the emotions in Emotion6 and IESN datasets are mainly determined by rich semantics, such as scenes and objects; (2) the proposed WMFSSL still achieves satisfying results without significant performance drop by removing some features, such as Attr in Abstract and GIST in IESN.

7 Conclusion

In this paper, we proposed to predict the probability distribution of image emotions, which can be viewed as an attempt to measure the subjectivity issue of emotion perceptions. We presented shared sparse learning as the learning model and extended it to multi-features settings, where both hand-crafted features and learning-based ones, are jointly explored. The optimal weights of different features that reflect their representation abilities are automatically learned. Experimental results on Abstract, Emotion6 and IESN datasets demonstrated the effectiveness of the proposed emotion distribution prediction method. For further studies, we plan to improve the computational efficiency of SSL and WMFSSL to tackle large-scale data. In addition, we will implement applications based on emotion distribution, such as image retrieval [36], [79], [80] and user opinion mining.

TABLE 5
Performance Comparison Between with and without PCA of WMFSSL on the Three Datasets Measured by SSD, KL, BC, R^2 ($\times 10^{-1}$) and the Standard Deviations ($\times 10^{-1}$), Where "-PCA" Represents WMFSSL without Dimension Reduction Using PCA

	SSD		KL		BC		R^2	
	-PCA	WMFSSL	-PCA	WMFSSL	-PCA	WMFSSL	-PCA	WMFSSL
Abstract	1.123 (0.054)	1.191 (0.060)	4.788 (0.257)	4.820 (0.209)	8.336 (0.089)	8.319 (0.078)	3.014 (0.375)	2.993 (0.467)
Emotion6	1.224 (0.040)	1.268 (0.076)	4.680 (0.013)	4.793 (0.097)	8.583 (0.021)	8.529 (0.059)	4.828 (0.030)	4.679 (0.172)
IESN	1.507 (0.015)	1.569 (0.014)	4.718 (0.043)	4.777 (0.016)	8.613 (0.034)	8.583 (0.015)	7.381 (0.392)	7.358 (0.382)

TABLE 6
Performance Contribution of Different Features in WMFSSL on the Three Datasets Measured by SSD, KL, BC, R^2 ($\times 10^{-1}$) and the Standard Deviations ($\times 10^{-1}$), Where "-GIST", "-Elem", "-Attr", "-Prin", "-ANP", and "-CNN" Represent WMFSSL without Using GIST, Elem, Attr, Prin, ANP, and CNN, Respectively

		-GIST	-Elem	-Attr	-Prin	-ANP	-CNN	WMFSSL
Abstract	SSD	1.198 (0.056)	1.244 (0.056)	1.189 (0.069)	1.255 (0.054)	1.225 (0.065)	1.259 (0.065)	1.191 (0.060)
	KL	4.887 (0.201)	5.047 (0.189)	4.819 (0.226)	5.050 (0.222)	4.975 (0.225)	5.106 (0.237)	4.820 (0.209)
	BC	8.261 (0.076)	8.242 (0.075)	8.323 (0.077)	8.201 (0.816)	8.269 (0.079)	8.192 (0.086)	8.319 (0.078)
	R^2	2.903 (0.392)	2.856 (0.418)	2.998 (0.466)	2.807 (0.477)	2.888 (0.431)	2.776 (0.444)	2.993 (0.467)
Emotion6	SSD	1.272 (0.075)	1.295 (0.067)	1.278 (0.070)	1.309 (0.087)	1.318 (0.091)	1.323 (0.076)	1.268 (0.076)
	KL	4.793 (0.077)	4.859 (0.085)	4.805 (0.082)	4.909 (0.154)	4.928 (0.135)	5.007 (0.091)	4.793 (0.097)
	BC	8.536 (0.049)	8.500 (0.047)	8.528 (0.061)	8.509 (0.070)	8.496 (0.073)	8.471 (0.050)	8.529 (0.059)
	R^2	4.645 (0.080)	4.605 (0.230)	4.654 (0.171)	4.558 (0.152)	4.507 (0.217)	4.465 (0.128)	4.679 (0.172)
IESN	SSD	1.571 (0.021)	1.575 (0.024)	1.594 (0.022)	1.623 (0.027)	1.633 (0.019)	1.658 (0.017)	1.569 (0.014)
	KL	4.795 (0.071)	4.796 (0.047)	4.806 (0.029)	4.840 (0.051)	4.869 (0.043)	4.899 (0.026)	4.777 (0.016)
	BC	8.585 (0.016)	8.581 (0.024)	8.565 (0.021)	8.527 (0.020)	8.505 (0.012)	8.457 (0.020)	8.583 (0.015)
	R^2	7.361 (0.395)	7.359 (0.312)	7.326 (0.369)	7.288 (0.387)	7.253 (0.386)	7.210 (0.372)	7.358 (0.382)

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Nos. 61701273, 61571269, 61671267, 61472103, 61772158, 61332016, 61620106009, U1636214), the Project Funded by China Postdoctoral Science Foundation (No. 2017M610897), the Royal Society Newton Mobility Grant (No. IE150997), and the National Key R&D Program of China (Grant No. 2017YFC011300). The authors would also like to thank the anonymous reviewers for their insightful comments to help us improve the paper.

REFERENCES

[1] 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 *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 367–376.

[2] H. Lin, J. Jia, L. Nie, G. Shen, and T.-S. Chua, "What does social media say about your stress?" in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3775–3781.

[3] W.-N. Wang, Y.-l. Yu, and S.-M. Jiang, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2006, pp. 3534–3539.

[4] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 101–104.

[5] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 83–92.

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

[7] 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 Int. Conf. Multimedia*, 2012, pp. 229–238.

[8] 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 *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 857–860.

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

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

[11] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 860–868.

[12] S. Zhao, H. Yao, X. Jiang, and X. Sun, "Predicting discrete probability distribution of image emotions," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 2459–2463.

[13] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T.-S. Chua, "Predicting personalized emotion perceptions of social images," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1385–1394.

[14] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 308–314.

[15] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multi-task shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, Mar. 2017.

[16] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 224–230.

[17] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3266–3272.

[18] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affective Comput.*, 2016, doi: 10.1109/TAFFC.2016.2628787.

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

[20] S. Zhao, "Image emotion computing," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1435–1439.

[21] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 3869–3872.

[22] C. Chen, J. Huang, L. He, and H. Li, "Preconditioning for accelerated iteratively reweighted least squares in structured sparsity reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2713–2720.

[23] S. Zhao, G. Ding, Y. Gao, and J. Han, "Approximating discrete probability distribution of image emotions by multi-modal features fusion," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 466–4675.

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

[25] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *Proc. ACM Int. Workshop Issues Sentiment Discovery Opinion Mining*, 2013, Art. no. 10.

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

[27] 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 Res. Methods*, vol. 37, no. 4, pp. 626–630, 2005.

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

[29] 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 Trans. Affective Comput.*, vol. 3, no. 2, pp. 132–144, Apr.-Jun. 2012.

[30] 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 *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 306–312.

[31] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 634–638.

[32] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3595–3601.

[33] H. Schlosberg, "Three dimensions of emotion." *Psychological Rev.*, vol. 61, no. 2, 1954, Art. no. 81.

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

[35] X. Alameda-Pineda, E. Ricci, Y. Yan, and N. Sebe, "Recognizing emotions from abstract paintings using non-linear matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5240–5248.

[36] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 1025–1028.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

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

[40] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Comput. Surveys*, vol. 49, no. 2, 2016, Art. no. 28.

[41] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

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

[43] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, 2012, Art. no. 40.

[44] A. Roda, S. Canazza, and G. De Poli, "Clustering affective qualities of classical music: Beyond the valence-arousal plane," *IEEE Trans. Affective Comput.*, vol. 5, no. 4, pp. 364–376, Oct.-Dec. 2014.

[45] S. Zhao, H. Yao, F. Wang, X. Jiang, and W. Zhang, "Emotion based image musicalization," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2014, pp. 1–6.

[46] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Modeling the affective content of music with a gaussian mixture model," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 56–68, Jan.-Mar. 2015.

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

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

[49] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 43–55, Jan.-Mar. 2015.

[50] S. Wang and Q. Ji, "Video affective content analysis: A survey of state of the art methods," *IEEE Trans. Affective Comput.*, vol. 6, no. 4, pp. 410–430, Oct.-Dec. 2015.

[51] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Jan.-Mar. 2012.

[52] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Trans. Affective Comput.*, vol. 7, no. 1, pp. 17–28, Jan.-Mar. 2016.

[53] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010.

[54] Y. Yang, P. Cui, W. Zhu, and S. Yang, "User interest and social influence based emotion prediction for individuals," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 785–788.

[55] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1008–1017.

[56] M. Carney, P. Cunningham, J. Dowling, and C. Lee, "Predicting probability distributions for surf height using an ensemble of mixture density networks," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 113–120.

[57] H. Liu, Z. Hu, D. Zhou, and H. Tian, "Cumulative probability distribution model for evaluating user behavior prediction algorithms," in *Proc. IEEE Int. Conf. Soc. Comput.*, 2013, pp. 385–390.

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

[59] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[60] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

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

[62] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[63] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 19, pp. 4–19, 2014.

[64] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[65] Y. Gao, Y. Zhen, H. Li, and T.-S. Chua, "Filtering of brand-related microblogs using social-smooth multiview embedding," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 2115–2126, Oct. 2016.

[66] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Cost-optimized microblog distribution over geo-distributed data centers: Insights from cross-media analysis," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, 2017, Art. no. 40.

[67] S. Zhao, Y. Gao, G. Ding, and T.-S. Chua, "Real-time multimedia social event detection in microblog," *IEEE Trans. Cybern.*, 2017, doi: 10.1109/TCYB.2017.2762344.

[68] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-d object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.

[69] S. Zhao, H. Yao, Y. Zhang, Y. Wang, and S. Liu, "View-based 3d object retrieval via multi-modal graph learning," *Signal Process.*, vol. 112, pp. 110–118, 2015.

[70] F. Wang, S. Qi, G. Gao, S. Zhao, and X. Wang, "Logo information recognition in large-scale social media data," *Multimedia Syst.*, vol. 22, no. 1, pp. 63–73, 2016.

[71] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1601–1608.

[72] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.

[73] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach Learn.*, 2011, pp. 689–696.

[74] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 471–478.

[75] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognit.*, vol. 38, no. 12, pp. 2437–2448, 2005.

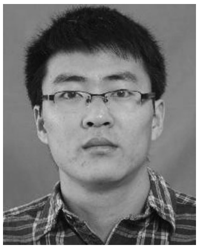
[76] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Fully automatic face normalization and single sample face recognition in unconstrained environments," *Expert Syst. Appl.*, vol. 47, pp. 23–34, 2016.

[77] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2751–2758.

[78] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[79] Y. Guo, G. Ding, L. Liu, J. Han, and L. Shao, "Learning to hash with optimized anchor embedding for scalable retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1344–1354, Mar. 2017.

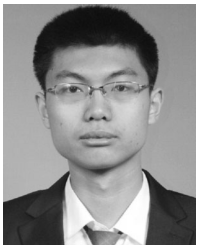
[80] Y. Guo, G. Ding, and J. Han, "Robust quantization for general similarity search," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 949–963, Feb. 2018.



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

Guiguang Ding received the PhD degree in electronic engineering from the University of Xidian, Xian, China. He is an associate professor with the School of Software, Tsinghua University, Beijing, China. His current research interests include the area of multimedia information retrieval and mining, with specific focus on visual object recognition, automatic semantic annotation, image coding and representation, and social media recommendation. He has published more than 40 research papers in international conferences and journals and applied for 18 Patent Rights in China.

Yue Gao (SM'14) received the BS degree from Harbin Institute of Technology, Harbin, China, and the ME and PhD degrees from Tsinghua University, Beijing, China. He is a senior member of the IEEE.



Xin Zhao received the bachelor's degree from the School of Software, Tsinghua University, in 2014. He is working toward the PhD degree in the School of Software, Tsinghua University. His current research interests include multimedia information retrieval and mining, in particular visual object retrieval, image caption, content-based multimedia indexing, and image classification.

Youbao Tang received the BSc, MSc, and PhD degrees in computer science from Harbin Institute of Technology, Harbin, China, in 2009, 2011, and 2016, respectively. He is a postdoc researcher with the National Institutes of Health. His research interests include image processing, computer vision, biometrics, and medical image analysis.

Jungong Han was with Civolution technology (a combining synergy of Philips CI and Thomson STS) from 2012 to 2015, a research staff with the Centre for Mathematics and Computer Science from 2010 to 2012, and a researcher with the Technical University of Eindhoven, The Netherlands from 2005 to 2010. He is currently a senior lecturer with the School of Computing & Communications, Lancaster University, United Kingdom.

Hongxun Yao received the BS and MS degrees in computer science from Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and 1990, respectively, and the PhD degree in computer science from Harbin Institute of Technology, Harbin, 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 more than 200 scientific papers. Her research interests include computer vision, pattern recognition, multimedia computing, and human–computer interaction technology. She received both the Honor Title of the New Century Excellent Talent in China and the Enjoy Special Government Allowances Expert in Heilongjiang, China.

Qingming Huang received the bachelor's degree in computer science and the PhD degree in computer engineering from the Harbin Institute of Technology, China, in 1988 and in 1994, respectively. He is currently a professor with the University of Chinese Academy of Sciences and an adjunct research professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include multimedia video analysis, image processing, computer vision, and pattern recognition. He has published more than 300 academic papers in prestigious international journals including the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Circuits and Systems for Video Technology* and top-level conferences, such as ACM Multimedia, ICCV, CVPR, IJCAI, and VLDB. He is an associate editor of Acta Automatica Sinica, and the reviewer of various international journals including the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Circuits and Systems for Video Technology*, and the *IEEE Transactions on Image Processing*. He has served the general chair, program chair, track chair, and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, and PSIVT. He is a fellow of the IEEE.