Learning Visual Emotion Distributions via Multi-Modal Features Fusion

Sicheng Zhao†, Guiguang Ding†, Yue Gao†, Jungong Han‡
†School of Software, Tsinghua University, Beijing 100084, China
‡School of Computing & Communications, Lancaster University, UK
schzhao@gmail.com, {dinggg,gaoyue}@tsinghua.edu.cn, jungonghan77@gmail.com

ABSTRACT

Current image emotion recognition works mainly classified the images into one dominant emotion category, or regressed the images with average dimension values by assuming that the emotions perceived among different viewers highly accord with each other. However, due to the influence of various personal and situational factors, such as culture background and social interactions, different viewers may react totally different from the emotional perspective to the same image. In this paper, we propose to formulate the image emotion recognition task as a probability distribution learning problem. Motivated by the fact that image emotions can be conveyed through different visual features, such as aesthetics and semantics, we present a novel framework by fusing multi-modal features to tackle this problem. In detail, weighted multi-modal conditional probability neural network (WMMCPNN) is designed as the learning model to associate the visual features with emotion probabilities. By jointly exploring the complementarity and learning the optimal combination coefficients of different modality features, WMMCPNN could effectively utilize the representation ability of each uni-modal feature. We conduct extensive experiments on three publicly available benchmarks and the results demonstrate that the proposed method significantly outperforms the state-of-the-art approaches for emotion distribution prediction.

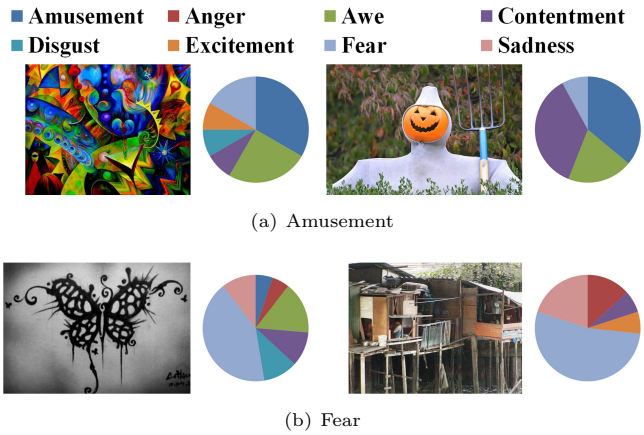
CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; • **Applied computing** → *Arts and humanities*;

KEYWORDS

Discrete probability distribution, image emotion, distribution learning, feature fusion, multi-modal conditional probability neural network

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10...$15.00
DOI: http://dx.doi.org/10.1145/3123266.3130858



(a) Amusement

(b) Fear

Figure 1: Affective image classification vs. emotion distribution prediction. Traditional affective image classification target classifying images into one dominant emotion category, such as (a) *Amusement* and (b) *Fear*, while emotion distribution prediction aims to predict the probabilities of different emotion categories, such as the pie chart on the right of each image. Besides, the image emotions may be conveyed through different visual features. The emotions of the left abstract paintings in each group are mainly determined by image aesthetics, such as *strong and weak color contrast*. The emotions of the right natural images mainly correspond to semantic concepts, such as *cute dummy* and *shabby house*.

1 INTRODUCTION

Being able to convey rich semantics, images have been widely used in people's daily lives to record their activities, express their opinions and share their experiences. Recently, lots of research attention has been paid to affective analysis of visual content for its broad applications [5], ranging from human-computer interaction to anomaly forewarning. The task of recognizing the emotions induced by visual content is often referred to as image emotion recognition (IER) [12, 35], which is often composed of three steps: human annotation collection, visual feature extraction and mapping learning between visual features and perceived emotions [34].

Similar to other visual recognition problems, one main challenge for IER is affective gap [35], i.e., the inconsistency between visual features and affective states. In the last

decade, researchers have designed various hand-crafted or learning-based features to bridge this gap. Assuming that different viewers react similarly to the same image, current IER methods mainly assign the image with the dominant emotion category (DEC). This task can be performed as a traditional single-label learning (SLL) problem.

However, labeling the visual emotions in ground-truth generation is in fact highly inconsistent. That is, different viewers may have totally different emotional reactions to the same image, which is caused by a variety of personal and situational factors, such as the cultural background, personality and social context [12, 21, 35, 36, 38, 39]. This phenomenon is called subjective perception problem, the categorial situation of which is shown in Figure 1. It is clear that despite with the same DEC, the two images of each group differ dramatically in terms of emotion variations. Therefore, we can conclude that just predicting the dominant emotion category is insufficient for this highly subjective variable.

To tackle the subjectivity issue, we can conduct two kinds of IER tasks [36, 38]: for each viewer, we can predict personalized emotion perceptions; for each image, we can assign multiple emotion labels. For the latter one, we can employ multi-label learning (MLL) methods [33], which associates one instance with multiple class labels. However, the importance or extent of different labels is in fact unequal. In such cases, label distribution learning would make more sense, which aims to learn the degree to which each label describes the instance [8], including discrete probability distribution (DPD) [21, 34, 39] and continuous probability distribution (CPD) [37]. Shared sparse learning (SSL) [39], support vector regression (SVR) and convolutional neural network regression (CNNR) [21] were employed to learn the mapping from visual features to emotion probabilities. Another state-of-the-art method is conditional probability neural network (CPNN) [9]. However, only uni-modal visual feature was modelled, which is obviously insufficient, since image emotions can be conveyed through different visual features from low-level to high-level [40], as shown in Figure 1. In addition, the SVR and CNNR approaches cannot guarantee that the predicted probability is non-negative; the signless integers used as label representation in CPNN make no sense when adding two emotion labels or subtracting one from another. Weighted multi-modal shared sparse learning (WMMSSL) [34] is the first work on fusing multi-modal features for emotion distribution prediction. But the sparse coefficients have to be learned each time, making it infeasible in practical applications.

In this paper, we propose a novel general framework to learn the DPD of image emotions from multi-modal visual features. Specifically, we extend CPNN into multi-modal settings and propose weighted multi-modal CPNN for emotion distribution learning. The framework of the proposed method is shown in Figure 2. In the emotion space, we count the human annotation numbers of different emotion categories and normalize the annotations to obtain the ground-truth DPDs for the training images. In the visual space, we extract multi-modal features from the images and use PCA to

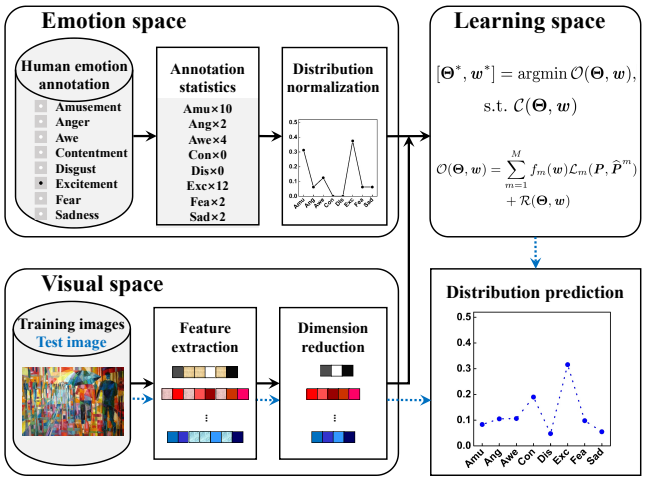


Figure 2: The framework of the proposed method for DPD prediction of image emotions by fusing multi-modal visual features. The black solid and blue dash arrowed lines indicate the operations in the training and testing stages, respectively.

reduce the feature dimensions [16, 35]. The visual features and ground-truth DPDs of the training images are input to the learning space to learn the model parameters. For a given test image, the extracted visual features are directly input to the learned model to predict the DPD. We conduct extensive experiments on Abstract [16], Emotion6 [21] and Image-Emotion-Social-Net (IESN) [36, 38] datasets that are labeled with distribution information to validate the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 reviews related work. The problem definition of emotion distribution learning is formalized in Section 3. We present the learning method and extracted features in Section 4. Experimental settings, results and analysis are given in Section 5, followed by the conclusion and future work in Section 6.

2 RELATED WORK

Image emotion recognition. As an interesting research topic for decades, IER has attracted active attention from both the academic and industrial communities for years. To represent emotions, there are typically two kinds of emotion models: categorical emotion states (CES) [6, 17] and dimensional emotion space (DES) [24]. CES methods directly map emotions to one of a few basic categories, while DES methods employ 3-D or 2-D space to represent emotions. CES model is straightforward for users to understand and label, while DES model is more descriptive. Accordingly, we can perform different tasks, including affective image classification, regression and retrieval [36, 38]. As the most traditional and popular research task, affective image classification aims to assign a dominant emotion category to an image. In this paper, we also represent image emotions based on CES models. But

we aim to predict the DPD of image emotions instead of focusing on the traditional single dominant emotion category.

Like many other vision tasks, the performance of IER greatly depends on feature extraction. To bridge the affective gap in the early years, researchers mainly designed different levels of hand-crafted features, including low-level color and texture [16], shape [15], mid-level principles-of-art [35] and high-level adjective noun pairs [2, 5, 28]. More recently, with the great success of convolutional nerual network (CNN) in many computer vision tasks [4], CNN has also been directly employed in IER [1, 25, 32]. In this paper, we extract hand-crafted features together with CNN ones, and jointly combine them for emotion distribution learning.

To map features to emotions, various machine learning methods have been employed, such as Naive Bayes [16], SVM or SVR [15, 35], multi-graph learning [40], sparse learning [37], matrix completion [1] and hypergraph learning [36, 38].

Label distribution learning. In many real-world machine learning applications, a single most likely value is often insufficient to describe a target variable. For example, in biology, the fungi genes over a period of time may yield different gene expression levels on a series of time points. The single level alone predicted by traditional SLL makes little sense. MLL [33] studies the problem where one instance is associated with a number of class labels, which could assign different expression levels for the time period, ignoring the degree to which each label describes the instance. But what really matters is the overall expression distribution over the whole time period. In such cases, the learning task becomes a LDL problem [8], which aims to predict the probability distribution for that variable [3], such as surf height [3], user behavior [14], spike events [22] and facial ages [9]. As demonstrated in [36, 38], the emotions that are evoked in viewers by an image are in fact subjective, which indicates that predicting the probability distribution instead of the dominant category is of more importance [39].

Generally, we can formalize the distribution prediction task as a regression problem. For CES, the task aims to predict the discrete probability of different emotion categories, the sum of which is equal to 1 [21, 39]. For DES, the task usually turns to predicting the parameters of specified continuous probability distribution, the form of which should be firstly decided, such as Gaussian distribution [37]. Geng *et al.* proposed CPNN to predict label distributions [9]. Further, existing learning methods in computer vision can be improved to deal with LDL problem through different strategies, such as problem transformation, algorithm adaption, and specialized algorithms design [8]. In this paper, we extend the uni-model CPNN into multi-modal settings by jointly exploring the representation ability of different features.

Multi-modal learning. With the wide-spread use of multimedia acquisition techniques and social media platforms, we might have multi-modal data to represent the same target [10], either from different sources [31] or with multiple features [40]. Typically, different modal data usually represent different aspects of the target, such as the global and local ortanizations. Therefore, jointly combining them may promisingly improve the performance [10, 26, 27]. Besides the traditional early fusion and late fusion methods, many other multi-modal fusion strategies have been proposed, such as multi-graph learning [29], hypergraph learning [7, 41], multi-modal deep learning [19] and multi-modal sparse learning [34]. However, to the best of our knowledge, fusing multi-modal features for probability distribution learning of visual emotions has been rarely explored.

3 PROBLEM DEFINITION

Our goal is to predict the DPD of image emotions when multi-modal features are available. Suppose we have L emotion categories $e_1, e_2, \cdots, e_L$ and N training images $I_1, I_2, \cdots, I_N$. The mth modal features of the N training images are $\boldsymbol{X}^m = [\boldsymbol{x}_1^m, \boldsymbol{x}_2^m, \cdots, \boldsymbol{x}_N^m]$ and the feature dimension is d_m ($m = 1, 2, \cdots, M$). Let $\boldsymbol{p}_n = [p_{n1}, \cdots, p_{nl}, \cdots, p_{nL}]^{\mathrm{T}}$ denote the emotion distribution of the image I_n, where p_{nl} represents the probability that image I_n conveys emotion e_l ($n = 1, 2, \cdots, N, l = 1, 2, \cdots, L$). For each image I_n, we have $p_{nl} \geq 0$ and $\sum_{l=1}^{L} p_{nl} = 1$. Suppose the parameters of the emotion distribution model p for the mth modality is $\boldsymbol{\theta}^m$ and the predicted emotion distribution turns to $\widehat{\boldsymbol{p}}_n^m = p(\boldsymbol{e}|\boldsymbol{x}^m; \boldsymbol{\theta}^m)$. Let $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_N]$ and $\widehat{\boldsymbol{P}}^m = [\widehat{\boldsymbol{p}}_1^m, \widehat{\boldsymbol{p}}_2^m, \cdots, \widehat{\boldsymbol{p}}_N^m]$ denote the ground-truth emotion distribution labels and predicted labels with the mth modality feature. Let $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \cdots, \boldsymbol{\theta}^M\}$, then the emotion distribution task aims to minimize the following objective function:

$$\begin{aligned} &\mathcal{O}(\boldsymbol{\Theta}, \boldsymbol{w}) = \sum_{m=1}^{M} f_m(\boldsymbol{w}) \mathcal{L}_m(\boldsymbol{P}, \widehat{\boldsymbol{P}}^m) + \mathcal{R}(\boldsymbol{\Theta}, \boldsymbol{w}), \\ &\text{s.t. } \mathcal{C}(\boldsymbol{\Theta}, \boldsymbol{w}), \end{aligned} \tag{1}$$

where $\mathcal{L}_m(\boldsymbol{P}, \widehat{\boldsymbol{P}}^m)$ is the loss function between two distribution sets for the mth modality, $f_m(\boldsymbol{w})$ is the combination coefficients of different loss functions, $\mathcal{R}(\boldsymbol{\Theta}, \boldsymbol{w})$ is the regularization term, and $\mathcal{C}(\boldsymbol{\Theta}, \boldsymbol{w})$ is the constraint.

Let $\mathcal{D}$ denote the distance measurement between two distributions, then one typical loss function form is

$$\mathcal{L}_m(\boldsymbol{P}, \widehat{\boldsymbol{P}}^m) = \sum_{n=1}^{N} \mathcal{D}(\boldsymbol{p}_n, \widehat{\boldsymbol{p}}_n^m). \tag{2}$$

Similar to [8, 9], we employ the Kullback-Leibler (KL) divergence as the distance measurement, and the loss function utilized in this paper is defined as

$$\mathcal{L}_m(\boldsymbol{P}, \widehat{\boldsymbol{P}}^m) = \sum_{n=1}^{N} KL(\boldsymbol{p}_n || \widehat{\boldsymbol{p}}_n^m) = \sum_{n=1}^{N} \sum_{l=1}^{L} p_{nl} \ln \frac{p_{nl}}{p(e_l|\boldsymbol{x}_n^m; \boldsymbol{\theta}^m)}. \tag{3}$$

Suppose J is a test image, its M modal features are $\boldsymbol{y}^1, \boldsymbol{y}^2, \cdots, \boldsymbol{y}^M$. Let $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2, \cdots, \boldsymbol{y}^M\}$. Once the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{w}$ are learned, the emotion distribution $\widehat{\boldsymbol{p}} = [\widehat{p}_1, \cdots, \widehat{p}_l, \cdots, \widehat{p}_L]^{\mathrm{T}}$, where $\widehat{p}_l = p(c_l|\boldsymbol{Y})$ for test image J can be predicted by

$$\widehat{\boldsymbol{p}} = \sum_{m=1}^{M} w_m p(\boldsymbol{e}|\boldsymbol{y}^m; \boldsymbol{\theta}^m). \tag{4}$$

4 PROPOSED METHOD

In practice, multi-modal visual features can be extracted to represent images [40]. Different features reflect different aspects of the images. For example, visual aesthetics may represent abstract paintings well, while semantic concepts can better describe natural scenes. Previous emotion distribution prediction methods mainly employed uni-modal features. CNNR [21] is based on CNN features, while SSL [39] and CPNN [9] compared different uni-modal features. Jointly combining multi-modal features to utilize the strengths from different aspects may improve the performance. SSL [39] and CPNN [9] can simply adopt early or late fusion to handle multi-modal features without considering the latent correlations between different features. WMMSSL [34] employs the constraint of joint sparsity across different features, which is the first work of modelling multi-modal features for discrete emotion distribution prediction. However, WMMSSL has to learn the sparse coefficients for each test image, which is time-consuming and unpractical in real applications. In this paper, we propose a novel method, named weighted multi-modal conditional probability neural network (WMMCPNN), for distribution learning of visual emotions by extending the CPNN method into multi-modal settings. Once the parameters are learned from the training data, the emotion distributions of the test images can be predicted in real time. The detailed model and the extracted emotion features are described in this section.

4.1 Weighted Multi-Modal Conditional Probability Neural Network

WMMCPNN aims to minimize the linear combination of the loss function in Eq. (3) with a sparse constraint on the combination coefficients, that is

$$[\boldsymbol{\Theta}^*, \boldsymbol{w}^*] = \operatorname{argmin} \sum_{m=1}^{M} w_m \sum_{n=1}^{N} \sum_{l=1}^{L} p_{nl} \ln \frac{p_{nl}}{p(e_l|\boldsymbol{x}_n^m; \boldsymbol{\theta}^m)} + \alpha||\boldsymbol{w}||_2^2, \quad \text{s.t. } \boldsymbol{w} \geq 0, \|\boldsymbol{w}\|_1 = 1, \tag{5}$$

where $p(e_l|\boldsymbol{x}_n^m; \boldsymbol{\theta}^m)$ is modelled by CPNN [9], a three layer neural network, as follows

$$p(e|\boldsymbol{x}; \boldsymbol{\theta}) = \exp(g(\boldsymbol{x}, e; \boldsymbol{\theta}) + b(\boldsymbol{x}; \boldsymbol{\theta})), \tag{6}$$

by taking both $\boldsymbol{x}$ and e as input. For convenience, we drop the superscript m and subscript n. Bias $b(\boldsymbol{x}; \boldsymbol{\theta})$ is defined as

$$b(\boldsymbol{x}; \boldsymbol{\theta}) = -\ln \left(\sum_{l=1}^{L} \exp(g(\boldsymbol{x}, e_l; \boldsymbol{\theta})) \right), \tag{7}$$

which is used to ensure that the predicted $p(e_l|\boldsymbol{x}; \boldsymbol{\theta}), (l = 1, 2, \cdots, L)$ is a probability distribution. The net activation of the output unit $g(\boldsymbol{x}, e; \boldsymbol{\theta})$ is

$$g(\boldsymbol{x}, e; \boldsymbol{\theta}) = \sum_{u=1}^{N_2} \theta_{31u} \left(\sum_{v=0}^{N_{11}} \theta_{2uv} x_v + \sum_{v=N_{11}+1}^{N_2} \theta_{2uv} \boldsymbol{h}(e) \right), \tag{8}$$

where $\mathcal{G}$ is the sigmoid activation function, N_t is the number of units on the tth layer, $x_0 \equiv 1$, $\boldsymbol{h}(e)$ is the transformed input of original input e to make the operation meaningful, and θ_{tuv} is the weight of the uth unit on the tth layer associated with the output of the vth unit on the $(t-1)$th layer.

We iteratively update $\boldsymbol{\Theta}$ or $\boldsymbol{w}$ when fixing the other to solve the dual-optimization problem in Eq. (5).

(1) Updating $\boldsymbol{\Theta}$ when fixing $\boldsymbol{w}$

When $\boldsymbol{w}$ is fixed, the optimization of each $\boldsymbol{\theta}^m$ in Eq. (5) is independent, and the corresponding target function to minimize turns to

$$\begin{aligned} \mathcal{O}(\boldsymbol{\theta}) &= -\sum_{n=1}^{N} \sum_{l=1}^{L} p_{nl} \ln \frac{p_{nl}}{p(e_l|\boldsymbol{x}_n; \boldsymbol{\theta})} \\ &= -\sum_{n=1}^{N} \sum_{l=1}^{L} p_{nl} (g(\boldsymbol{x}_n, e_l; \boldsymbol{\theta}) + b(\boldsymbol{x}_n; \boldsymbol{\theta})), \end{aligned} \tag{9}$$

where the superscript m is omitted for convenience. The gradient of Eq. (9) with respect to $\boldsymbol{\theta}$ is

$$\frac{\partial \mathcal{O}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{n=1}^{N} \sum_{l=1}^{L} p_{nl} \left(\frac{\partial g(\boldsymbol{x}_n, e_l; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\partial b(\boldsymbol{x}_n; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right), \tag{10}$$

where

$$\frac{\partial b(\boldsymbol{x}_n; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\frac{\sum_{l=1}^{L} \left(\exp(g(\boldsymbol{x}, e_l; \boldsymbol{\theta})) \times \frac{\partial g(\boldsymbol{x}_n, e_l; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)}{\sum_{l=1}^{L} \exp(g(\boldsymbol{x}, e_l; \boldsymbol{\theta}))}, \tag{11}$$

and the partial derivative of $g(\boldsymbol{x}, e_l; \boldsymbol{\theta})$ can be calculated by back propagation [18] as follows,

$$\frac{\partial g(\boldsymbol{x}, e_l; \boldsymbol{\theta})}{\partial \theta_{tuv}} = z_{(t-1)v}^n \delta_{tu}^n, \tag{12}$$

where $z_{(t-1)v}^n$ is the output of the vth unit on the $(t-1)$th layer, and δ_{tu}^n is the partial derivative of $g(\boldsymbol{x}, e_l; \boldsymbol{\theta})$ with respect to the net activation of the uth unit on the tth layer I_{tu}^n. For the output layer, i.e. when $t = 3$,

$$\delta_{31}^n = \frac{\partial g(\boldsymbol{x}, e_l; \boldsymbol{\theta})}{\partial I_{31}^n} = 1. \tag{13}$$

For the hidden layer, i.e. when $t = 2$,

$$\delta_{2u}^n = \frac{\partial g(\boldsymbol{x}, e_l; \boldsymbol{\theta})}{\partial I_{2u}^n} = \mathcal{G}'(I_{2u}^n) \delta_{31}^n \theta_{31u} = \mathcal{G}'(I_{2u}^n) \theta_{31u}. \tag{14}$$

After $\frac{\partial \mathcal{O}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is obtained, the weights can be updated by the RPROP algorithm [23].

(2) Updating $\boldsymbol{w}$ when fixing $\boldsymbol{\Theta}$

The optimization problem of Eq. (5) with respect to $\boldsymbol{w}$ when $\boldsymbol{\Theta}$ is fixed is a standard quadratic programming problem, which can be easily and efficiently solved by off-the-shelf quadratic optimization methods. The learning procedure is summarized in Algorithm 1.

When $M = 1$, the proposed WMMCPNN turns to a simplified version, named CPNN [9].

4.2 Extracted Emotion Features

As shown in [40], there are different types of features that may contribute to the perceptions of image emotions. We extract various emotion features to jointly explore the representation power, including hand-crafted ones of different levels and learning-based ones.

Algorithm 1: Procedure for weighted multi-modal conditional probability neural network learning

Input: Training examples $\{\boldsymbol{X}^1, \boldsymbol{X}^2, \cdots, \boldsymbol{X}^M, \boldsymbol{P}\}$, test feature $\boldsymbol{Y}$, max-epochs E, error threshold τ_1, τ_2, regularization coefficients α

Output: Predicted emotion distribution $\widehat{\boldsymbol{p}}$ for $\boldsymbol{Y}$

Initialization: $\boldsymbol{\theta}^{m(0)} \leftarrow \mathbf{1}(m = 1, 2, \cdots, M)$, $\boldsymbol{w}^{(0)} \leftarrow \mathbf{1}/M$;

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

  `/* Updating Θ when fixing w */`

  for $m \leftarrow 1$ ***to*** M **do**

    Compute $\dfrac{\partial \mathcal{O}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ by Eq. (10);

    Optimize $\boldsymbol{\theta}^{m(e)}$ by updating $\boldsymbol{\theta}^{m(e-1)}$ with RPPOP [23];

  end

  `/* Updating w when fixing Θ */`

  Optimize $\boldsymbol{w}$ by

$$\boldsymbol{w}^{(e)} \leftarrow \operatorname{argmin} \sum_{m=1}^{M} w_m \sum_{n=1}^{N} \sum_{l=1}^{L} p_{nl} \ln \frac{p_{nl}}{p(e_l|\boldsymbol{x}_n^m; \boldsymbol{\theta}^{m(e)})}, + \alpha||\boldsymbol{w}||_2^2, \quad \text{s.t. } \boldsymbol{w} \geq 0, \|\boldsymbol{w}\|_1 = 1,$$

  if $\|\boldsymbol{\theta}^{m(e)} - \boldsymbol{\theta}^{m(e-1)}\|_2 < \tau_1 (m = 1, 2, \cdots, M)$ *and* $\|\boldsymbol{w}^{(e)} - \boldsymbol{w}^{(e-1)}\|_2 < \tau_2$ **then**

    break;

  end

end

return $\widehat{\boldsymbol{p}} = \sum_{m=1}^{M} w_m p(\boldsymbol{e}|\boldsymbol{y}^m; \boldsymbol{\theta}^{m(e)})$ by Eq. (6).

Though suffering from the difficulty of interpretation and weak link to emotions [40], two classes of low-level hand-crafted features are first extracted for their global descriptions of the overall image content, including generic GIST [20] and the features derived from elements-of-art (such as *color* and *texture*) [16]. Being more interpretable and having stronger link to emotions than low-level ones [35], mid-level features are recently widely used in various visual recognition tasks. Two classes of mid-level features are extracted, including attributes (such as *smoothing* and *open area*) [20] and features inspired from principles-of-art (such as *balance*, *contrast*, *harmony* and *variety*) [35]. High-level features are the detailed semantic contents contained in images, the conveyed emotions of which can be easily understood by humans via recognizing the semantics. Here we extract a set of concepts described by adjective noun pairs (ANPs, such as *beautiful flowers* and *cute boy*), detected by a large detector library SentiBank [2], which is trained on about 500k images downloaded from Flickr using various low-level features, including GIST, color histogram, *etc.* Further, we extract the deep features from the response of the fully connected layer (FC) 7 of the ImageNet-CNN [13], which is the final fully connected layer before producing the class predictions.

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.

Figure 3: Some image examples in Abstract (top), Emotion6 (middle) and IESN (bottom) datasets.

5 EXPERIMENTS

In this section, we first introduce the detailed experimental settings, including the three public datasets that contain DPD information of visual emotions, compared baselines, evaluation metrics and implementation details. Then we evaluate the performance of the proposed emotion distribution learning method, report and analyze the results as compared to the state-of-the-art approaches.

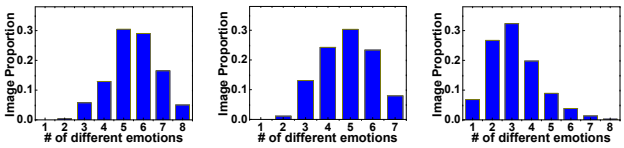
5.1 Experimental Settings

Datasets: The Abstract dataset [16] contains 279 abstract paintings which consist only of combinations of color and texture, without any recognizable objects. Based on the Mikels' emotion model [17], the images were peer rated in a web-survey by approximately 230 people into 8 emotion categories (see Figure 1), where each image was rated 14 times on average. Please note that only 228 images can be used for affective image classification [16], while all the images can be used for emotion distribution prediction.

The Emotion6 dataset [21] consists of 1,980 images collected from Flickr by using the emotion keywords and synonyms as search terms. Different from Abstract [16], the Ekman's 6 basic emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*) are used as the emotion categories. There are 330 images for each emotion category. Amazon Mechanical Turk (AMT) workers were invited to label the images into the 6 basic emotions and *neutral* to obtain the emotional responses. Each image was scored by 15 subjects.

The IESN dataset [36, 38] contains 1,012,901 images collected from Flickr using keywords based searching strategy [2, 30]. Instead of time-consuming labeling, the emotion information of the social images in IESN are automatically obtained from the corresponding text data. Similar to Abstract [16], the emotions are also classified into 8 categories. This dataset was initially designed for personalized emotion perception prediction. In this paper, we select 3,792 images in total for emotion distribution learning, each of which is assigned with more than 15 categorial emotion labels.

Some image examples in the three datasets are illustrated in Figure 3, the image styles of which differ a lot from each other. We can easily obtain the ground-truth emotion



(a) Abstract [16] (b) Emotion6 [21] (c) IESN [36, 38]

Figure 4: The distribution of images that are labeled with different emotion numbers. It is clear that the majority of images are labeled with at least two emotion categories, which demonstrates that the perceived emotions are truly subjective.

distribution on the 8 or 7 categories of each image by normalization, i.e., dividing the number of subjects who perceive each emotion category by the number of all emotion perceptions. For instance, suppose the perceived emotion numbers of an image by 40 subjects on the 8 emotion categories are $v = [6, 4, 8, 14, 3, 6, 4, 5]$, then the DPD can be obtained by $v/\sum(v) = [0.12, 0.08, 0.16, 0.28, 0.06, 0.12, 0.08, 0.1]$. Note that one subject may perceive more than one emotion from the same image. The image distributions with different emotion numbers are shown in Figure 4, which clearly show the subjectivity issue of emotion perceptions.

Baselines: To compare with the state-of-the-art approaches, we select shared sparse learning (SSL) [39], conditional probability neural network (CPNN) [9], convolutional neural network regression (CNNR) [21] and weighted multi-modal SSL [34] as baselines. The former three ones are based on uni-modal features, while the last one employs multi-modal features. Further, we implement early and late fusion for SSL and CPNN to deal with multi-modal features. Similar to [21], we pre-train the Caffe reference model [11] and fine-tune the CNN with our training set for CNNR.

Evaluation Metrics: Similar to [34], we employ the sum of squared difference (SSD) [39], the Kullback-Leibler divergence (KL)[1], the Bhattacharyya coefficient (BC)[2] and the coefficient of determination (R^2)[3] as evaluation metrics. $0 \leq SSD \leq 1$, $KL \geq 0$ and lower values represent better performances. BC ranges from 0 to 1 and larger value indicates better result. $0 \leq R^2 \leq 1$ and larger value indicates stronger linear relationship between two distributions.

Implementation Details: Similar to [34], 80%, 50% and 50% of images are randomly selected from the Abstract, Emotion6 and IESN datasets respectively as the training set and the rest constitute the testing set. Unless otherwise specified, $\alpha = 1000$ is adopted in experiment. Empirical analysis on parameter sensitivity is also conducted, which demonstrates that WMMCPNN has superior and stable performance with a wide range of parameter values on all the three datasets. $\boldsymbol{h}(e_l)$ is set as a binary vector to replace traditional signless integer e_l, i.e., $\boldsymbol{h}(e_l) = [b(e_1), \cdots, b(e_l), \cdots, b(e_L)]$, where $b(e_l) = 1$ and other binary values are 0. Since m is small enough as compared to the feature vectors, this specific design would not affect the computational complexity much. For fair comparison, we carefully tune the parameters of the baselines and report the best results. Further, we perform 20 runs and report the average results to remove the influence of any randomness.

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

5.2 Results and Discussion

On Uni-Modal Feature Based Methods

First, we conduct experiments to compare the performance of uni-modal feature based methods (i.e., SSL [39], CPNN [9] and CNNR [21]) and different visual features (i.e., GIST, Elem, Attr, Prin, ANP and CNN) for emotion distribution learning. Table 1, Table 2 and Table 3 present the performances of different emotion distribution learning methods measured by SSD, KL, BC, R^2 on Abstract, Emotion6 and IESN datasets, respectively. In the middle column, the best results of uni-modal feature based methods are highlighted in italic.

From the results, we can observe that: (1) the CNN features outperform the hand-crafted ones in general; the high-level and mid-level hand-crafted features have stronger discriminability than low-level ones, which are consistent with current methods on visual emotion analysis [32, 36, 38, 40]; (2) SSL and CPNN achieve comparable results on most features; (3) the CNNR method outperforms the other uni-modal feature based method in most cases, which demonstrates its effectiveness in emotion distribution learning [21]; (4) the adopted performance metrics SSD, KL, BC and R^2 are relatively consistent with each other for distribution measure; (5) for both SSL and CPNN, Elem and Prin even outperform ANP on Abstract and Emotion6 datasets; on Abstract dataset, the performance of Prin is superior to both ANP and CNN in CPNN method; this phenomenon results from the fact that the abstract paintings in Abstract do not contain recognizable objects [16], while the images in Emotion6 with apparent emotion related semantics are removed in dataset construction [21]; (6) SSD, KL and BC are with similar values across different datasets, while R^2 is much larger in IESN dataset, the emotion numbers of which are smaller (see Fig. 4) due to the influence of social and temporal factors.

On Different Feature Fusion Methods

Second, the performance comparison is conducted among different feature fusion methods, including the proposed WMMCPNN, early fusion and late fusion for CPNN and SSL, and WMMSSL. On the right column of Table 1, Table 2 and Table 3, we show the better fusion method for SSL and CPNN in bold, and highlight the best overall result in both italic and bold.

From the results, we have the following observations: (1) both the early fusion and late fusion methods for SSL as well as CPNN outperform uni-modal features; (2) the best fusion method for SSL depends on the datasets with late, early and early fusion methods achieving the best results

Table 1: Performance comparison between the proposed WMMCPNN method with the state-of-the-art approaches, including SSL [39], CPNN [9], CNNR [21] and WMMSSL [34], for discrete emotion distribution prediction on Abstract dataset [16] measured by SSD, KL, BC and R^2 ($\times 10^{-1}$).

	SSL						CPNN						CNNR	SSL		CPNN		WMMSSL	WMMCPNN
	GIST	Elem	Attr	Prin	ANP	CNN	GIST	Elem	Attr	Prin	ANP	CNN	CNN	Early	Late	Early	Late	Fusion	Fusion
SSD	1.369	1.354	1.474	1.346	1.316	1.282	1.441	1.328	1.469	1.315	1.330	1.321	*1.244*	1.271	**1.241**	1.488	1.227	**1.191**	***1.088***
KL	5.525	5.439	6.070	5.421	5.475	5.225	5.627	5.381	5.909	5.436	5.501	5.495	*5.103*	5.126	**5.034**	6.041	5.013	**4.820**	***4.611***
BC	7.992	8.095	7.849	8.106	8.118	8.118	8.027	9.119	7.852	8.174	8.148	8.128	*8.173*	8.142	**8.210**	8.062	8.218	**8.319**	***8.442***
R^2	1.915	2.161	1.850	2.478	2.483	2.660	1.936	2.050	1.967	2.687	2.525	2.462	*2.796*	2.678	**2.817**	2.593	2.824	**2.993**	***3.067***

Table 2: Performance comparison between the proposed WMMCPNN method with the state-of-the-art approaches, including SSL [39], CPNN [9], CNNR [21] and WMMSSL [34], for discrete emotion distribution prediction on Emotion6 dataset [21] measured by SSD, KL, BC and R^2 ($\times 10^{-1}$).

	SSL						CPNN						CNNR	SSL		CPNN		WMMSSL	WMMCPNN
	GIST	Elem	Attr	Prin	ANP	CNN	GIST	Elem	Attr	Prin	ANP	CNN	CNN	Early	Late	Early	Late	Fusion	Fusion
SSD	2.043	1.828	1.984	1.806	1.794	1.427	2.066	1.786	1.952	1.728	1.707	1.561	*1.394*	**1.344**	1.402	1.452	**1.295**	1.268	***1.231***
KL	6.389	5.999	6.205	5.863	5.705	5.244	6.554	5.791	6.111	5.428	5.408	5.346	*4.846*	**4.825**	5.064	5.138	**4.800**	4.793	***4.642***
BC	7.876	7.940	7.909	8.073	8.151	8.402	7.943	8.049	7.946	8.110	8.185	8.238	*8.437*	**8.484**	8.411	8.375	**8.493**	8.529	***8.631***
R^2	2.755	3.601	2.832	3.644	3.683	4.237	2.810	3.563	2.715	3.714	3.786	4.052	*4.434*	**4.533**	4.368	4.282	**4.602**	4.679	***4.745***

Table 3: Performance comparison between the proposed WMMCPNN method with the state-of-the-art approaches, including SSL [39], CPNN [9], CNNR [21] and WMMSSL [34], for discrete emotion distribution prediction on IESN dataset [36, 38] measured by SSD, KL, BC and R^2 ($\times 10^{-1}$).

	SSL						CPNN						CNNR	SSL		CPNN		WMMSSL	WMMCPNN
	GIST	Elem	Attr	Prin	ANP	CNN	GIST	Elem	Attr	Prin	ANP	CNN	CNN	Early	Late	Early	Late	Fusion	Fusion
SSD	1.928	1.854	1.863	1.852	1.728	1.719	1.913	1.880	1.895	1.847	1.733	1.714	*1.703*	**1.676**	1.706	1.706	**1.602**	1.569	***1.532***
KL	5.606	5.292	5.173	5.083	4.915	4.874	5.588	5.320	5.458	5.053	4.988	4.864	*4.828*	**4.802**	4.837	4.845	**4.792**	4.777	***4.696***
BC	7.910	8.258	8.319	8.385	8.425	8.515	8.142	8.455	8.303	8.443	8.475	8.506	*8.534*	**8.542**	8.524	8.475	**8.565**	8.583	***8.704***
R^2	6.828	7.043	7.119	7.201	7.221	7.232	7.029	7.123	7.165	7.221	7.283	*7.306*	*7.306*	**7.314**	7.265	7.249	**7.321**	7.357	***7.684***

on Abstract, Emotion and IESN datasets, respectively; (3) late fusion of CPNN is superior to early fusion, which is consistent across all the three datasets; (4) using the best fusion method, CPNN performs a bit better than SSL, and both of them outperform CNNR; (5) the proposed fusion method, namely WMMCPNN, achieves the best results on all the three datasets, as compared to the baselines, including the latest WMMSSL [34], which demonstrates the effectiveness of WMMCPNN in learning visual emotion distributions by jointly fusing multi-modal features.

Specifically, the performance gains of CPNN with late fusion method over the best uni-modal features measured by SSD, KL, BC, R^2 are 6.69%, 7.79%, 0.52%, 5.08% on Abstract, 17.01%, 10.21%, 3.10%, 13.56% on Emotion6 and 6.52%, 1.47%, 0.68%, 0.20% on IESN datasets, respectively. Compared with the best results of SSL, CPNN, CNNR and WMMSSL, WMMCPNN achieves the KL performance gains of 12.32%, 11.35%, 12.51%, 8.61% on Abstract, 12.23%, 4.99%, 11.72%, 2.96% on Emotion6 and 10.12%, 4.35%, 10.01%, 2.31% on IESN datasets, respectively. These results demonstrate that the proposed WMMSSL achieves significant performance improvements as compared to the state-of-the-art approaches for emotion distribution learning.

Figure 5 shows the predicted emotion distributions on some images from different datasets. We can see that the proposed WMMCPNN generates more similar emotion distributions to the ground truth than the baselines.

On Parameter Sensitivity

In the proposed WMMCPNN, we have one model parameter α to control the relative importance between the loss function and regularization term. In this subsection we report how sensitive WMMCPNN is to the parameter.

The influence of the parameter α on WMMCPNN is validated, with results shown in Figure 6. From these curves, we can find that: (1) the influence of α is different across different datasets; more stable performance is obtained on Emotion6 dataset than Abstract and IESN datasets; (2) on Abstract dataset, when increasing α, the performance firstly becomes better and then turns to be worse, meaning that there exists the best α; (3) on IESN dataset, the performance becomes worse when α reaches 1000; (4) the metric R^2 is more robust than the rest metrics. These results reveal the robustness of the proposed WMMCPNN method for visual emotion distribution learning.

6 CONCLUSION

In this paper, we tackled the subjectivity challenge of visual emotion perceptions by learning the probability distributions instead of single dominant emotion category. A novel distribution learning framework is presented by fusing multi-modal

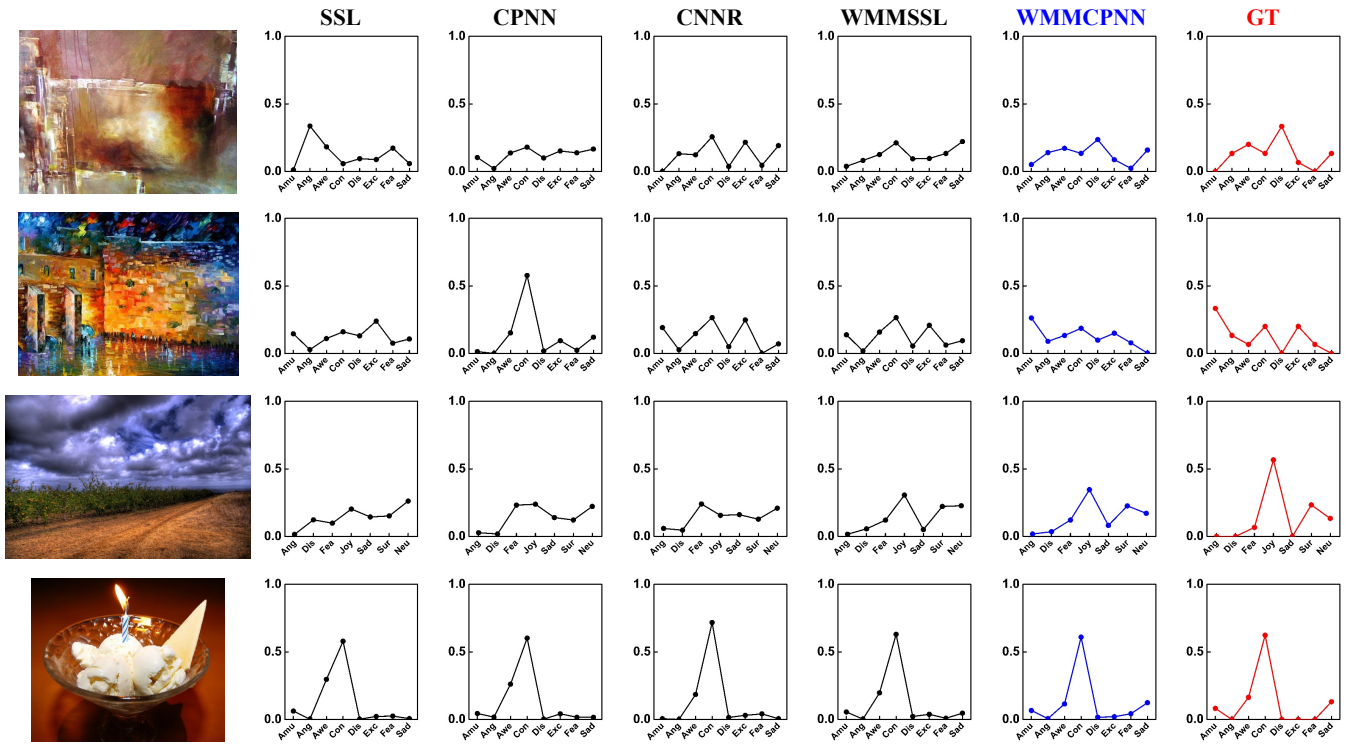


Figure 5: Visualization of predicted emotion distributions using the proposed WMMCPNN and the state-of-the-art approaches (SSL [39] and CPNN [9] with the related best uni-modal feature, CNNR [21] and WMMSSL [34]). Original images and the corresponding ground truth distributions ('GT') are shown in the first and last columns of each group, respectively.

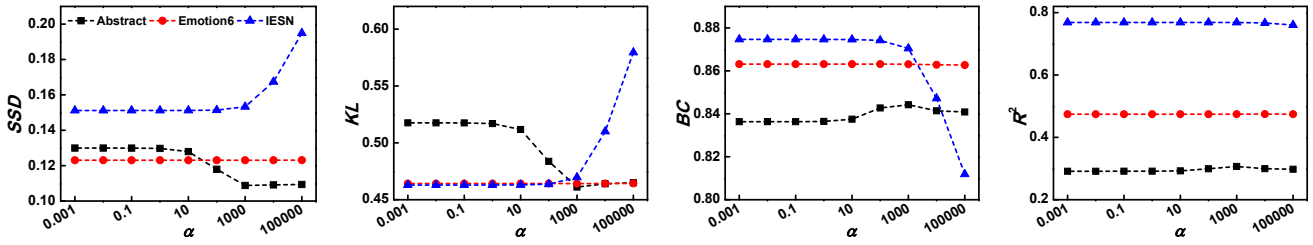


Figure 6: The influence of the regularization coefficient α in the proposed WMMCPNN method.

features. We extended the state-of-the-art conditional probability neural network (CPNN) into multi-modal settings and proposed weighted multi-modal CPNN to jointly explore the representation ability of different modality features. The model parameters and the optimal combination coefficients of different features are jointly and automatically learned. Both hand-crafted features and learning-based ones are fused on three publicly available benchmarks, including Abstract, Emotion6 and IESN. The experimental results demonstrated that the proposed method is superior to the contrastive baselines. For further studies, we plan to extend the three-layer neural network to deep CNN architecture. How to design and implement effective and efficient loss function and regularization term in the presented framework is worth studying.

7 ACKNOWLEDGEMENTS

This research was supported by the Project Funded by China Postdoctoral Science Foundation (No. 2017M610897), the National Natural Science Foundation of China (No. 61571269, 61671267) and the Royal Society Newton Mobility Grant (IE150997). The authors would also like to thank the anonymous reviewers for their insightful comments to help us improve the paper.

REFERENCES

[1] Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe. 2016. Recognizing emotions from abstract paintings using non-linear matrix completion. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5240–5248.

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM International Conference on Multimedia*. 223–232.

[3] Michael Carney, Pádraig Cunningham, Jim Dowling, and Ciaran Lee. 2005. Predicting probability distributions for surf height using an ensemble of mixture density networks. In *International Conference on Machine Learning*. 113–120.

[4] Minghai Chen, Guiguang Ding, Sicheng Zhao, Hui Chen, Qiang Liu, and Jungong Han. 2017. Reference Based LSTM for Image Captioning. In *AAAI Conference on Artificial Intelligence*. 3981–3987.

[5] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. 2014. Object-based visual sentiment concept analysis and application. In *ACM International Conference on Multimedia*. 367–376.

[6] Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion* 6, 3-4 (1992), 169–200.

[7] Yue Gao, Sicheng Zhao, Yang Yang, and Tat-Seng Chua. 2015. Multimedia Social Event Detection in Microblog.. In *International Conference on Multimedia Modeling*. 269–281.

[8] Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1734–1748.

[9] Xin Geng, Chao Yin, and Zhi-Hua Zhou. 2013. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 10 (2013), 2401–2412.

[10] Alex Pappachen James and Belur V Dasarathy. 2014. Medical image fusion: A survey of the state of the art. *Information Fusion* 19 (2014), 4–19.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*. 675–678.

[12] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[14] Haifeng Liu, Zheng Hu, Dian Zhou, and Hui Tian. 2013. Cumulative Probability Distribution Model for Evaluating User Behavior Prediction Algorithms. In *IEEE International Conference on Social Computing*. 385–390.

[15] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. 2012. On shape and the computability of emotions. In *ACM International Conference on Multimedia*. 229–238.

[16] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*. 83–92.

[17] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. 2005. Emotional category data on images from the International Affective Picture System. *Behavior Research Methods* 37, 4 (2005), 626–630.

[18] Dharmendra S Modha and Yeshaiahu Fainman. 1994. A learning law for density estimation. *IEEE Transactions on Neural Networks* 5, 3 (1994), 519–523.

[19] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *International Conference on Machine Learning*. 689–696.

[20] Genevieve Patterson and James Hays. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2751–2758.

[21] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. 2015. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 860–868.

[22] Gordon Pipa, Sonja Grün, and Carl van Vreeswijk. 2013. Impact of Spike Train Autostructure on Probability Distribution of Joint Spike Events. *Neural Computation* 25, 5 (2013), 1123–1163.

[23] Martin Riedmiller and Heinrich Braun. 1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks*. 586–591.

[24] Harold Schlosberg. 1954. Three dimensions of emotion. *Psychological Review* 61, 2 (1954), 81.

[25] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen. 2016. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *IEEE International Conference on Multimedia and Expo*. 1–6.

[26] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291.

[27] Johannes Wagner, Elisabeth Andre, Florian Lingenfelser, and Jonghwa Kim. 2011. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing* 2, 4 (2011), 206–218.

[28] Jingwen Wang, Jianlong Fu, Yong Xu, and Tao Mei. 2016. Beyond Object Recognition: Visual Sentiment Analysis with Deep Coupled Adjective and Noun Neural Networks. In *International Joint Conference on Artificial Intelligence*. 626–630.

[29] Meng Wang, Xian-Sheng Hua, Richang Hong, Jinhui Tang, Guo-Jun Qi, and Yan Song. 2009. Unified video annotation via multigraph learning. *IEEE Transactions on Circuits and Systems for Video Technology* 19, 5 (2009), 733–746.

[30] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. 2014. How Do Your Friends on Social Media Disclose Your Emotions?. In *AAAI Conference on Artificial Intelligence*. 306–312.

[31] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. 2016. Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. In *ACM International Conference on Multimedia*. 1008–1017.

[32] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI Conference on Artificial Intelligence*. 308–314.

[33] Min-Ling Zhang and Lei Wu. 2015. Lift: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 1 (2015), 107–120.

[34] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. 2017. Approximating Discrete Probability Distribution of Image Emotions by Multi-Modal Features Fusion. In *International Joint Conference on Artificial Intelligence*.

[35] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *ACM International Conference on Multimedia*. 47–56.

[36] Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. 2016. Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing* (2016).

[37] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. 2017. Continuous Probability Distribution Prediction of Image Emotions via Multi-Task Shared Sparse Regression. *IEEE Transactions on Multimedia* 19, 3 (2017), 632–645.

[38] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. 2016. Predicting personalized emotion perceptions of social images. In *ACM International Conference on Multimedia*. 1385–1394.

[39] Sicheng Zhao, Hongxun Yao, Xiaolei Jiang, and Xiaoshuai Sun. 2015. Predicting discrete probability distribution of image emotions. In *IEEE International Conference on Image Processing*. 2459–2463.

[40] Sicheng Zhao, Hongxun Yao, You Yang, and Yanhao Zhang. 2014. Affective image retrieval via multi-graph learning. In *ACM International Conference on Multimedia*. 1025–1028.

[41] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. 2006. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*. 1601–1608.