

Received August 13, 2019, accepted August 26, 2019, date of publication September 5, 2019, date of current version September 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939681

Predicting Image Emotion Distribution by Learning Labels' Correlation

YANGYU FAN¹, HANSEN YANG¹, ZUHE LI², AND SHU LIU³

¹School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China

²School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China

³School of Computer Science and Engineering, Central South University, Changsha 410083, China

Corresponding author: Zuhe Li (zuheli@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702462 and Grant 61902435, in part by the Scientific and Technological Project of Henan Province under Grant 182102210607 and Grant 192102210108, and in part by the Hunan Provincial Natural Science Foundation of China under Grant 2019JJ50808.

ABSTRACT Image emotion analysis attracts considerable attention with the increasing demanding of opinion mining in social networks. Emotion evoked by an image is always ambiguous for emotion's subjectivity. Different from previous researches on image emotion classification, Label Distribution Learning framework which assigns a set of labels with degree value to an instance, describes emotions more explicitly. However, in our study, we find that some labels have co-occurrence relation with others and all the labels together appear some structural forms. To make use of these relations as complementary information to the holistic distribution of labels, we analysis the correlations among emotion labels and then propose a method based on Structural Learning framework, which learns a mapping from images to the distribution labels with correlations. On the other hand, images usually contain some emotion-unrelated contents, to extract features that can represent image emotion at utmost, we propose a cropping method to select the emotional region from the images with the help of Fully Convolutional Networks. Extensive experiments on two widely used datasets show the advantages of our methods.

INDEX TERMS Emotional region, fully convolutional neural networks, label distribution learning, structural Learning.

I. INTRODUCTION

Social networks become more and more popular in recent years and they are important medium for people to share their opinions, emotions and daily lives. Huge number of texts, pictures and videos are propagated among different users [1], [2]. These create vast requirements of processing and understanding the contents on social networks. Nowadays, analysis of image emotion begins to play an important role in opinion mining and monitoring, consumption trend and market predicting [3] etc. However, image emotion analysing is a demanding task.

One of the challenges is that the subjective perception problem. One can have multiple emotions towards an image with different intensities and different people have individualized compound emotions towards one image [4]. Most existing studies take image emotion prediction as a classification

problem in which the dominant emotion category is selected as the prediction result and the outputs give a certain emotion [3], [4]. But a certain label ignores explicit difference on emotion intensity. Thus classification doesn't coincide with subjective perception problem. Recently, Label Distribution Learning (LDL) algorithms were proposed [5]–[9] successively. In LDL, each label is associated with a description degree value which describes how likely the instance belongs to the label and all the values on each label form a distribution of an instance. The target of LDL is learning a mapping to minimizing the loss between input and output distributions. Figure 1 shows several images and their distribution labels. However, we notice that some emotion are usually accompanied by others, such as 'Surprise' and 'Joy' often appear simultaneously and so do 'Anger' and 'Fear'. But some emotions are unlikely to have both high description degree values, for example, 'Disgust' and 'Joy' etc, as the colorized bars in Figure 1. Based on these phenomena, we assume that correlations exist among emotion labels. Consequently,

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.

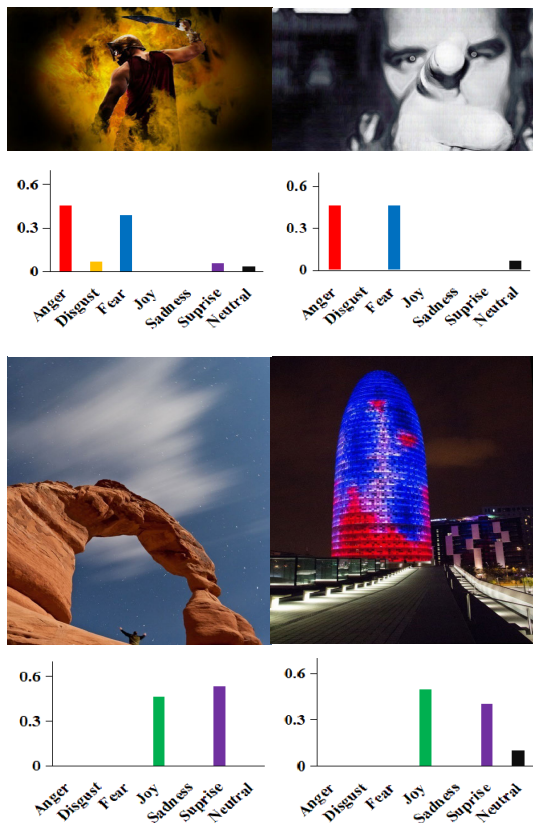


FIGURE 1. Emotional images and their distribution labels below. Some label often has high description degree with another simultaneously or contrarily.

we take analysis on the on image emotion distribution only learned the mapping from inputs to outputs but correlations among each member of output are ignored. Structural Learning [10] is a framework of the learning problem associated with outputs' correlation and it plays important role in the fields of multi-class classification, sequence tagging, object tracking [10], etc. Out of the correlations of emotion labels, in this work, we recast LDL of image emotion as Structural Learning problem. To the best of our knowledge, this is the first work which analyzes and learns emotion correlation in the field of image emotion analysis.

On the other hand, we study emotional local features of images. Recently, some studies are starting to notice the features from regions. Sun *et al.* [11] detected objects in image and they predicted the emotion value of each object by a special model. An algorithm was designed to synthesize all the objects with high emotion value to compute the sentiment of the images. Besides the objects, some other areas also could induce emotion. Moreover, the emotion value predicting model can lead into errors. Li *et al.* [12] segmented images into both objects and contexts, and each segmentation result was considered to make contribution to emotion. Rao *et al.* [13] firstly segmented images into 1, 4, 16 blocks corresponding to 3-levels. For each block, deep features were extracted and the final representation of image is the average of 21 representation vectors of the blocks.

In these studies, the local features are from the whole image, and the contribution of each part is assumed to be equal, but in fact, contents located at different areas gain different importance, features unrelated to emotion could reduce prediction accuracy. Instead of detecting objects or segmenting images by semantics, we propose to crop the region that influence emotion most in images, with the help of Fully Convolutional Neural Networks (FCN) [14]. FCN achieved success in Semantic Segmentation field, where each pixel is classified into one category in label space. We use the main architecture of FCN to generate a heatmap the same size with the input image by following Peng's [15] work, each pixel of the heatmap represents the probability of which each pixel of image evokes emotion. After normalized, the heatmap is represented by values between 0 to 1 in pixel-wise. They defined the heatmap as Emotion Stimuli Map (ESM), we crop the emotional region according to ESM. Both middle-level features (MLFs) and deep features are extracted from the cropped region to train Structural Learning models. Experiments conducted on two widely used image emotion datasets show the advantage of our methods.

The main contributions of our work can be summarized as follows:

- We take analyses on image emotion datasets and reveal the interdependency of the emotion labels.
- We recast image emotion distribution learning as Structural Learning, and construct the Structural Learning model by utilizing the labels' interdependency.
- We propose an image cropping method based on FCN model to get features that can represent image emotion most.

The rest of the paper is organized as follows: Firstly we will briefly review the most relevant contents in Section II, and then explain our proposed method in detail in Section III. Experiments are conducted to validate our approach on emotion label distribution dataset in Section IV. At last we will draw a conclusion of this work.

II. RELATED WORKS

A. LABEL DISTRIBUTION LEARNING

Most widely used image emotion datasets are built for Single Label Learning (SLL), for example, Twitter [16] and DeepSenti [17] are labeled with the highest number of statistics by different voters, so the prediction task is dividing the instance to a certain category. But in reality, a person usually has multiple emotions towards one image, and also, different people could have different emotions for one image. Multi-Label Learning (MLL) [18], which assigns logical values (usually 0 or 1) to multiple labels, could reflect subject perception to a certain extent, but fails to illustrate the degree to which each label describes the image. To explicitly describe the discrepancy between different labels for one instance, a general learning paradigm called Label Distribution learning [5], has been proposed, which assigns multiple label to one instance, and the description degrees

are affiliated with each label. Existing LDL algorithms are mainly based on three frameworks: 'Problem Transformation' (PT), 'Algorithm Adaptation' (AA), 'Specialized Algorithm' (SA) [5]. For 'Problem Transformation', LDL problem was transformed into weighted single-label problem, Bayes classifier and SVM model that can output the probabilities (regarded as description values) for each label are adopted, they were called PT-Bayes and PT-SVM for short. For 'Algorithm Adaptation', AA- k NN [8] outputted the label distribution for an instance by computing the mean of k nearest neighbor instances. AA-BP [9] used a three layers back-propagation (BP) neural network, the number of input units was the dimension of feature vector and each last layer unit outputted a description degree of each label. For 'Specialized Algorithm', SA-IIS algorithm was designed for facial age estimation [9], label distributions were generated from single label by the Gaussian distribution, it optimized the Kullback-Leibler loss function of maximum entropy model to predict distribution, SA-BFGS [6] was based on SA-IIS and improved the optimization method. Condition Probability Neural Network (CPNN) [9] was based on a three layers neural networks, both features and labels were input into the neural networks, it outputted probabilities of each label.

Some other algorithms dealing with LDL were proposed. Peng *et al.* [19] built Emotion6 database with distribution labels, and proposed an approach called CNR which contained 7 Support Vector Regression (SVR) [20], each SVR was based on a Convolutional Neural Networks (CNN) [21] for each emotion category, but the number of output nodes is changed to 1 to predict a description value. They also transformed original image emotion distribution to a target distribution. Zhao *et al.* [22] proposed to formalize the emotion distribution prediction as a shared sparse learning problem, and integrated several hand-crafted features to train their model. Afsheen *et al.* [23] trained a group of SVRs, each SVR can output a value to represent the description degree corresponding to one emotion, and they used Linear Admixture Model to study the relation between high level concepts and emotion distributions. Most of the algorithms for LDL problem are based on probability generation or regression models, only the mapping from features to labels are learned. However, the local correlation among some labels can be good complementary information for learning [24]–[27]. Structural learning has been studied intensively in some approaching fields. Earlier, researchers used Structural SVM in multi-class classification [28], natural language parsing [29] etc. Recently, Michal *et al.* [30] predicted apparent ages by Structural SVM, some correlations exist in the rates of age, e.g., when one rater gives a certain rating, it is highly possible that the other raters will give the ratings similar to the former rating. Wu *et al.* [31] studied web aesthetics problem, they captured the correlations between mean score and standard deviation of the aesthetics ratings by constructing combined features, and then optimize Structural SVM by learning the combined features. As a more general form of MLL, label dependency in LDL is reflected in

continuous numerical (description degree values to one label), rather than only assigning a logical value to one label.

B. REGION OF INTEREST

As mentioned before, what evokes people's emotion is usually in local of the images [4], [11], but how to localize them remains a tough task. Li *et al.* [3] detected objects and selected those represent emotion most and designed an algorithm to compute image emotion by these objects. Rao *et al.* [30] segmented images into squares and combined all the deep representation of squares to train a classifier. Li *et al.* [12] segmented image into several parts and extracted features from each part to represent the whole image.

By adopting 1980 images in Emotion6, Peng *et al.* [15] built a dataset and named it as EmotionROI. The groundtruths were got by asking the subjects from Amazon Mechanical Turk (AMT) to draw 15 rectangles enclosing the part of the image that influences emotion most. They assumed the influence of each pixel on evoked emotion is proportional to the times of pixel being covered by rectangles. Then the groundtruths are normalized to 1. After trained on EmotionROI, FCN model can output a heatmap in which areas with different colors represent the degree of influence on emotion and they defined the heatmap as ESM. Besides where to select features that are concerned with emotion, we also care about what kind of features to extract in our studies. Deep features from pre-trained CNN [21] show powerful ability on general visual tasks. Firstly we use deep features to train our Structural Learning model. Specially, MLFs from Sentibank gain great performance on image emotion analysing and exceed some hand-crafted and low-level features [31]. Therefore MLFs are compared with deep features in later experiments.

III. METHODS

The framework of our proposed methods is in Figure 2. It consists of three parts: (1) Predicting ESM by fine-tuned FCN, (2) Generating bounding box by ESM and cropping images, (3) Structural Learning model is trained and predicts emotion distributions.

A. EMOTIONAL REGION

FCN is adopted to generate ESM. FCN has been shown excellent performance in Semantic Segmentation. It is an end-to-end framework that can predict in the same resolution as the input image in pixel-wise dense. Following Peng's work [15], FCN with single stream and 32-pixel-prediction stride version is adopted, and it's based on the relative simple AlexNet architecture [34] for the size of EmotionROI database is not so large. When predicting ESM, the task is not segmenting image into several categories but predicting the probability of evoking emotion on each pixel, so the last fully connected layer of FCN is replaced by only one output. Original softmax loss is replaced by Euclidean loss so that emotion probability of pixels can be predicted. Example ESMs can be

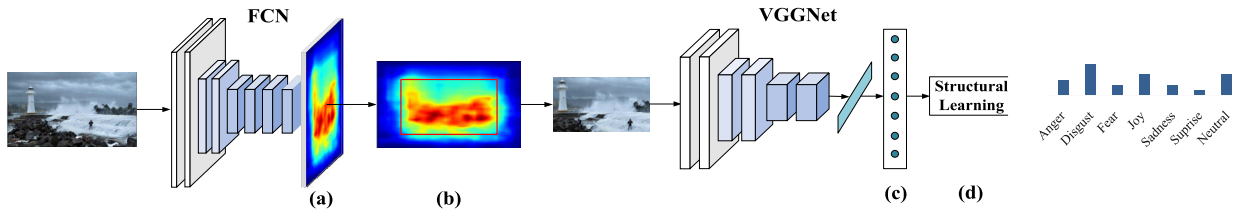


FIGURE 2. Pipeline of our proposed approach. a) The output of FCN is ESM. b) The smallest rectangle bounding box is returned to cover the pixels that evoke emotion with high probability. c) Deep features are extracted from emotional region. d) Structural Learning model.

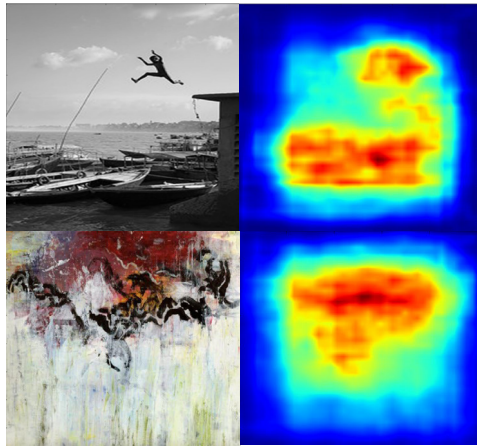


FIGURE 3. Examples of ESMs (right column) and corresponding original images (left column).

seen in Figure 3, different colors represent varied pixel-wise possibility of output in ESM, we attempt to crop the original image so as to locate the part which can represent emotion of image most. To get the region that preserves the most emotional part of the image while occupies minimum area, we set a probability threshold λ and for each image in both of the training and testing sets, and return the smallest rectangle bounding box which covers all pixels that have equal or greater possibility than λ , as shown in Figure 2 b). The original images and ESMs have the same size, we can crop the original image with the bounding box aforementioned in the same position with ESM, as shown in Figure 2 c), stuffs in bounding box represent emotion of the original images with high probability. Then we resize the cropped region to 224×224 and feed it into feature extractor subsequently.

B. STRUCTURAL LEARNING

Our task is to generate a mapping from features space to label space, by considering correlations among labels, e.g., some labels would have high description degree simultaneously while others are in opposite manner, so if we fix each label in a ordinal way, the distribution appears a structured form.

Assuming that we have a training set $D = \{(x_1, d_1), (x_2, d_2), \dots, (x_N, d_N)\}$ with N instances, x_i is the i -th instance and $d_i = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \dots, d_{x_i}^{y_M}\}$ is the distribution label, $\{y_1, y_2, \dots, y_M\}$ is M emotion labels, $d_{x_i}^{y_j}$ denotes the description degree of the j -th emotion label and satisfies $\sum_j d_{x_i}^{y_j} = 1$, $j \in \{1, 2, \dots, M\}$.

Suppose X is the feature space and S is the distribution label space, LDL learns a mapping function $f: X \rightarrow S$ from D . By considering of the correlations among labels, we attempt to solve this by Structural Learning framework [29], [31], which aims to learn a discriminant function $\Phi: X \times S \rightarrow R$ with input-output pairs, Φ can be considered as a matching function. Ideally, the maximum of it is achieved at the most matching label distribution under a certain loss. Therefore the form of f is represented as:

$$f(x, w) = \operatorname{argmax}_{d \in S} \Phi(x, d; w) \quad (1)$$

In standard Structural Learning [29], [35], Φ is a linear combination of combined feature:

$$\Phi(x, d; w) = \langle w, \psi(x, d) \rangle \quad (2)$$

w is the linear combination coefficient to be found in model, $\psi(x, d)$ is the combined feature which captures not only the relation between inputs and outputs, but also the correlation among outputs, the specific form of $\psi(x, d)$ will be introduced later.

Since our target is assigning the maximum value of Φ to the most matched sample, in the process of training w , the margin between matched samples and mismatched ones should be maximized.

The learning process of f is minimizing the risk function as following [29]:

$$R(f) = \int_{X \times S} l(y, f(x)) dD(x, y) \quad (3)$$

where $D(x, y)$ is the joint distribution, and $l(y, f(x))$ is the prediction loss.

Algorithm 1 Update the Working Set (W) With (x_i, d_i) in the l -th Iteration

Input: $w_l, (x_i, d_i), W_l$,
 1: $\hat{d}_i = \operatorname{argmax}_{\hat{d}_i \in S} l(d_i, \hat{d}_i) + \langle w, \psi(x_i, \hat{d}_i) \rangle$
 2: **if** $\frac{1}{N} \sum_{i=1}^N l(d_i, \hat{d}_i) - \frac{1}{N} \sum_{i=1}^N \langle w, \Delta \psi_i(\hat{d}_i) \rangle > \xi + \varepsilon$
 then $W_{l+1} = W_l \cup \hat{d}_i$
 else $W_{l+1} = W_l$
 Output: W_{l+1}

Structural SVM [35] (Available on line: <http://svmlight.joachims.org/>):

$$\begin{aligned} \min_{\mathbf{w}, \xi \geq 0} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C\xi, \quad \forall (\mathbf{y}_1, \dots, \mathbf{y}_N) \in S \\ \text{s.t.} \quad & \frac{1}{N} \sum_{i=1}^N \langle \mathbf{w}, \Delta \psi(\mathbf{y}_i) \rangle \geq \frac{1}{N} \sum_{i=1}^N l(\mathbf{y}_i, \mathbf{d}_i) - \xi \end{aligned} \quad (4)$$

where $\Delta \psi(\mathbf{y}_i) = \psi(\mathbf{x}_i, \mathbf{d}_i) - \psi(\mathbf{x}_i, \mathbf{y}_i)$, and ξ is slack variable, C is penalty factor.

Euclidean distance is adopted as the loss function to compute the difference between labels distributions and predicted distributions:

$$l(\mathbf{y}_i, \mathbf{d}_i) = \sqrt{\sum_{j=1}^M |y_{x_i}^j - d_{x_i}^j|^2} \quad (5)$$

The key problem to solve in Eq. (5) is that the first constraints for each training sample is infinite. Following [33], we solve this problem by adopting Cutting-plane method which condense infinite constraints to finite ones called 'working set' which starts with a null set. In each iteration of Cutting-plane algorithm, $\mathbf{w}$ is computed by solving Eq. (5) with the current working set, then the most 'violated constraints' (according to the desired precision ε) are found by Algorithm 1 for (x_i, d_i) and are added to the work set.

Finally the constraint in Eq. (5) can be rewritten as following:

$$\forall (\mathbf{y}_1, \dots, \mathbf{y}_N) \in W$$

The definition of $\psi(\mathbf{x}, \mathbf{d})$ depends on the characteristic of the specific task, we formulate it by two parts:

$$\psi(\mathbf{x}, \mathbf{d}) = [\psi_1(\mathbf{x}, \mathbf{d})^T, \psi_2(\mathbf{d})^T]^T \quad (6)$$

The first part $\psi_1(\mathbf{x}, \mathbf{d})$ reflects the correlations among input features and output distributions. As our study is a general form of multiple-class learning, we define $\psi_1(\mathbf{x}, \mathbf{d})$ by referring [10], it is defined as following:

$$\psi_1(\mathbf{x}, \mathbf{d}) = \mathbf{x} \otimes \mathbf{d} = [x_1 d_x^{y_1}, \dots, x_p d_x^{y_M}]^T \quad (7)$$

p is the feature dimension and $\otimes$ is the tensor product, ψ_1 is the exhaustion of possible combinations of each dimension in features and labels.

To construct the discriminant function, quantificationally analysing is conducted in our study. Pearson correlation coefficient matrices of labels are given in Figure 4, we can see that the coefficients in different position are coincident with the observation mentioned in Section I previously. In Figure 4 a), 'Dis', 'Sad', 'Surp', 'Neu' are the abbreviations of 'Disgust', 'Sadness', 'Surprise', 'Neutral' in Emotion6 dataset, and in Figure 4 b), 'Amus', 'Cont', 'Dis', 'Excit' represent 'Amusement', 'Contentment', 'Disgust', 'Excitement' in Abstract dataset. In coefficient matrices, if the absolute value of the unit is equal or greater than 0.3, the two corresponding labels are deemed to be interdependent and those less than 0.3 are independent. To describes the correlations among

emotional labels, we compute all possible products of correlated labels' value without repeat and arrange all products into a vector:

$$\psi_2(\mathbf{d}) = [d_x^{y_{j_1}} d_x^{y_{j_2}}, \dots, d_x^{y_{j_k}} d_x^{y_{j_l}}]^T \quad (8)$$

where $d_x^{y_{j_k}} d_x^{y_{j_l}}$ is the product of correlated label's description degrees, j_k and j_l are serial number of the correlated labels and $1 \leq k < l \leq M$.

IV. EXPERIMENTS

A. DATASETS

We evaluate our proposed methods on two widely used emotion distribution datasets: Emotion6 [19] and Abstract [36].

The Abstract dataset contains 279 abstract paintings without concrete objects. These images were voted by about 230 people into 8 emotion categories i.e. anger, amusement, awe, contentment, disgust, excitement, fear and sadness. Each image was rated about 14 times on average. For each label of instance was rated, they can be used as emotion distribution dataset.

Peng *et al.* retrieved the Ekman's 6 basic emotions (anger, disgust, fear, joy, sadness, surprise) [37] words and their synonyms from Flickr, by adding a natural emotion category, totally 1980 images in Emotion6 dataset were collected, each emotion category has 330 images. They employed AMT workers to vote to each image and count the number of votes for each emotion category, each image was voted by 15 subjects, by dividing sum vote of 7 emotion categories, label distribution for an image was got. Dividing vote of each emotion by sum vote of 7 emotion categories, distribution of each image can be obtained. Assuming 15 subjects vote on 7 emotion categories, is $\{1, 0, 10, 20, 0, 7, 2\}$ respectively, so the label distribution is $\{0, 0.25, 0.5, 0, 0.175, 0.05\}$.

B. FEATURES

Borth *et al.* [33] constructed a large-scale Visual Sentiment Ontology (VSO) by using Adjective Noun Pairs (ANPs). They retrieved images with the same ANP from Flickr. For each ANP, the collected Images are used to train a concept detector with some hand-crafted features. Totally, 1200 detectors with high responses are reserved, named Sentibank (Available on line: <http://visual-sentiment-ontology.appspot.com/>). Given an image, Sentibank outputs a 1200 dimensional vector which can be seen as a MLF for the image. Meanwhile, deep features extracted from the last fully connected layer of VGG16 [38] are taken as representation of input image, we reduce it to 300 dimensional by Principle Component Analysis (PCA) [39].

C. SETTING

We compared our proposed algorithms with several state-of-the-art LDL methods: AA-kNN, AA-BP, PT-Bayes, PT-SVM, SA-IIS, SA-BFGS, CNNR. For AA-BP, the number of hidden-layer neurons is set to 50, and k for AA-kNN is set as 10. BFGS-LDL and IIS-LDL follow the advised settings in [5]. PT-Bayes and PT-SVM are with the linear kernel.

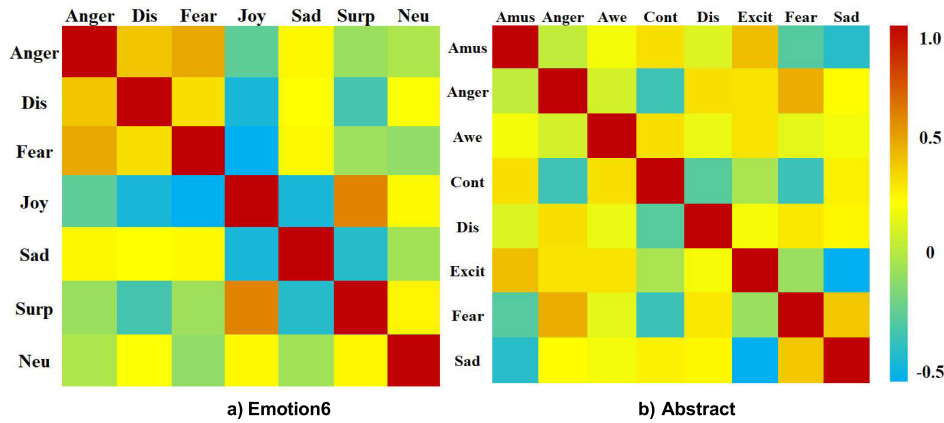


FIGURE 4. The correlation coefficients of emotion labels on different datasets.

In our Structural Learning model, C is set to 400, ε is 0.1 and ξ is initialized with 0. The CNN model in CNNR is pre-trained on ImageNet and fine-tuned with our training set. For all the methods we validate, 80% of the image are selected for training and the rest images are tested. For CPNN, both features and labels are regarded as input of a three layers neural networks, and it was especially designed for age estimation, the numerical labels are corresponding to the chronological ages, but emotion has no relationship with numbers, sending emotion labels into CPNN is meaningless, so we don't conduct experiments on it.

D. MEASUREMENT METRICS

To compare predicted distributions with ground truths. We adopt six metrics mentioned in [38], include Chebyshev distance, Clark distance, Canberra metric, Kullback-Leibler divergence, Cosine coefficient and Intersection similarity. The six metrics are introduced as following:

Chebyshev distance is the maximum value between two points in vector space. The Chebyshev distance (Cheb) between the ground truth and predicted distribution is determined by:

$$Cheb = \frac{1}{m} \sum_{i=1}^m \max_j \left| d_{x_i}^{y_j} - p(y_j | x'_i; \theta) \right|$$

The smaller Clark distance value indicates better performance. The Clark distance (Clark) between the ground truth and the predicted distribution is given as follows:

$$Clark = \frac{1}{m} \sum_{i=1}^m \sqrt{\sum_{j=1}^c \frac{(d_{x_i}^{y_j} - p(y_j | x'_i; \theta))^2}{(d_{x_i}^{y_j} + p(y_j | x'_i; \theta))^2}}$$

Variables in Canberra metric are assumed to be independent with each other, the correlations between variables are not considered. The Canberra distance (Canber) between the

ground truth and the predicted distribution is given as follows:

$$Canber = \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^c \frac{\left| d_{x_i}^{y_j} - p(y_j | x'_i; \theta) \right|}{d_{x_i}^{y_j} + p(y_j | x'_i; \theta)}$$

Cosine distance (Cosine) is similarity metric which measures the difference between two vector directions. When the directions of the two vectors are coincident, cosine gets the maximum value 1, and the cosine value is -1 when the directions of the two vectors are exactly opposite. It is given as follows:

$$Cosine = \frac{1}{m} \sum_{i=1}^m \frac{\sum_{j=1}^c d_{x_i}^{y_j} p(y_j | x'_i; \theta)}{\sqrt{\sum_{j=1}^c (d_{x_i}^{y_j})^2} \sqrt{\sum_{j=1}^c (p(y_j | x'_i; \theta))^2}}$$

Kullback-Leibler divergence represents the difference between two functions or probability distributions. The value of Kullback-Leibler divergence (KL) are 0 when two distributions are identical, it is given as follows:

$$KL = \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^c d_{x_i}^{y_j} \ln \frac{d_{x_i}^{y_j}}{p(y_j | x'_i; \theta)}$$

Since KL is not well defined when a description value is 0, we use a very small value 10^{-6} to approximate 0 description values.

Intersection (Interse) is a similarity metric which indicates the intersection of two vector sets, the larger intersection value indicates better performance. It is formulated as follows:

$$Interse = \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^c \min(d_{x_i}^{y_j}, p(y_j | x'_i; \theta))$$

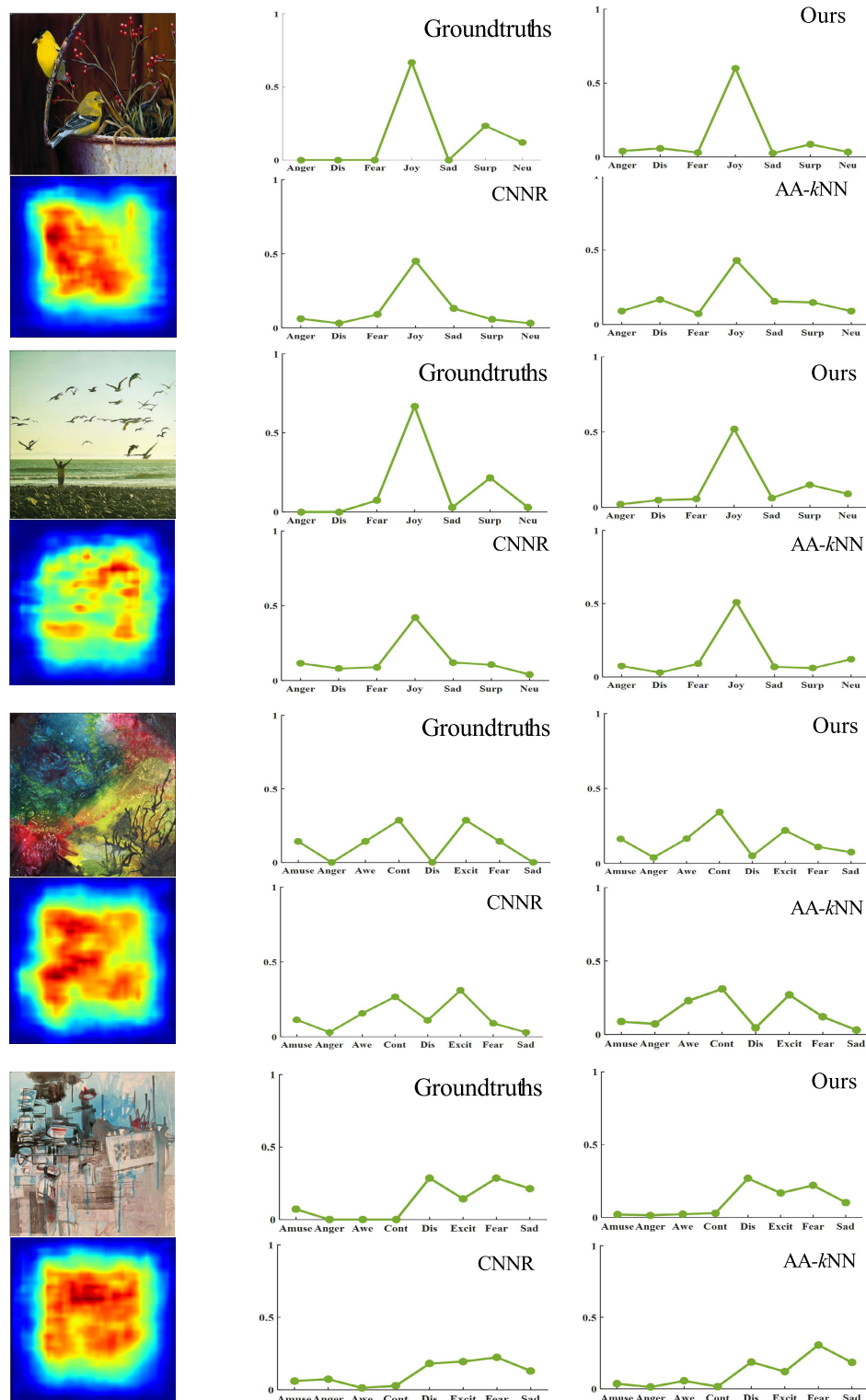


FIGURE 5. Some qualitative results of several comparable methods and ours.

V. RESULTS AND ANALYSIS

In this section, we show the results based the methods we propose. And firstly, we contrast the effects of dimension and the type of feature. Then, we launch analysis from experiments based on the baselines and our methods.

A. EFFECT OF FEATURES

Deep features are widely used in image emotion analysis and play key role in visual understanding [23], [41], we first use deep features to conduct experiments on two datasets. For original deep features are high-dimensional, we reduce them

TABLE 1. Performance comparison of different dimensions. '↓' indicates smaller values are better and '↑' indicate the larger the better. For each metric, the best performance is emphasized in boldface.

		240	260	280	300	320	340
Abstract	Cheb↓	0.2732	0.267	0.2590	0.2505	0.2561	0.2685
	Clark↓	1.8345	1.8299	1.8276	1.7729	1.7925	1.8071
	Canber↓	4.5152	4.2088	4.1975	4.2061	4.2132	4.2279
	KL↓	0.5251	0.5237	0.5195	0.5124	0.5211	0.5284
	Cosine↑	0.7492	0.7678	0.7721	0.7747	0.7705	0.7683
	Interse↑	0.5816	0.6212	0.6244	0.6267	0.6253	0.6208
Emotion6	Cheb↓	0.2930	0.2764	0.2684	0.2601	0.2662	0.2781
	Clark↓	1.7330	1.7086	1.7162	1.6738	1.7149	1.7785
	Canber↓	3.6359	3.5184	3.5125	3.5072	3.5247	3.5561
	KL↓	0.5413	0.5199	0.5226	0.4916	0.5125	0.5178
	Cosine↑	0.7076	0.7205	0.7911	0.7780	0.7947	0.7914
	Interse↑	0.5875	0.6049	0.6122	0.6204	0.6289	0.6054

TABLE 2. Performance comparison of different features.

	Abstract		Emotion6	
	Sentibank	Deep features	Sentibank	Deep features
Cheb↓	0.2578	0.2505	0.2643	0.2601
Clark↓	1.7857	1.7729	1.6981	1.6738
Canber↓	4.2265	4.2061	3.5216	3.5072
KL↓	0.5184	0.5124	0.5020	0.4916
Cosine↑	0.7688	0.7747	0.7498	0.7580
Interse↑	0.6317	0.6267	0.5902	0.6204

by PCA. To empirically determine the reduced dimension, we select different numbers of principle components, ranging from 240 to 340, repeated the experiments under the same other conditions. The variation of prediction performances versus the reduced dimension are shown in Table 1. Since the MLFs from Sentibank are related with emotion and proven to be good representation on visual sentiment tasks, we compare it with deep features to select more effective for consequent studies, the same dimension reduction experiments on MLFs are also conducted, 340-dimension gains optimum, contrast results with deep are in Table 2.

B. EFFECT OF ESM

In this section, according to the conclusion in Section A, we conduct experiments based on the deep features. Firstly, we validate the effects of different values of cropping threshold. Results are shown in Table 3. 0 value of λ is identical to no cropping and the cropped region is too small to cover the emotional region when the value of λ is close to 1. Finally, λ is set to 0.7. In Table 4, 'EDSL' represents our emotion distribution structural learning approach, 'CEDSL' represents both EDSL and cropping approaches are conducted. The best performance is highlighted in boldface and those exceed the baselines are in underline format. We can observe that prediction metrics with cropping exceed that from the holistic

images. It proves that human emotions are indeed influenced by the region of images, which reinforces the effectiveness of our cropping method. However, the improvements of cropping on Abstract are not so prominent compared with those on Emotion6, the possible reason for this phenomenon is that emotional contents in Emotion6 are distributed in a relative small region in general, so the cropping exclude more noisy contents, while the emotional elements occupy the majority part of the abstract paintings.

C. COMPARISONS WITH STATE-OF-THE-ART LDL ALGORITHMS

We compared our proposed algorithm with several state-of-the-art LDL algorithms, including PT-Bayes [5], PT-SVM [5], AA-kNN [8], AA-BP [9], SA-IIS [9], SA-BFGS [6] and CNNR [19]. Features are from VGGnet and reduced to 300-dimension. The experimental results are shown in Table 3. We can observe such following facts and take analyses correspondingly: (1) CNNR provides comparable prediction results, because for CNNR, the training process for multiple emotion categories are all under supervised, by minimizing the target function of each emotion, the optimal emotion distribution is achieved, but the disadvantage of CNNR is that each regressor may output a negative value, to construct the label distribution, the negative values are converted

TABLE 3. Effects of different cropping threshold.

Dataset	Abstract				Emotion6			
λ	0	0.4	0.7	0.9	0	0.4	0.7	0.9
Cheb↓	0.2505	0.2567	0.2411	0.5746	0.2601	0.2585	0.2447	0.5916
Clark↓	1.7729	1.8027	1.8232	2.9247	1.6838	1.6898	1.6113	2.8101
Canber↓	4.2061	4.1989	4.1816	5.9583	3.5472	3.4822	3.4702	6.1347
KL↓	0.5124	0.5073	0.5010	0.8598	0.4916	0.4817	0.4636	0.8392
Cosine↑	0.7747	0.7786	0.7804	0.5024	0.7580	0.7706	0.7939	0.4538
Interse↑	0.6267	0.6251	0.6293	0.4561	0.6204	0.6285	0.6514	0.4579

TABLE 4. Comparisons with state-of-the-art LDL algorithms.

Datasets	Criterion	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-IIS	SA-BFGS	CNNR	EDSL	CEDSL
Abstract	Cheb↓	0.3687	0.3215	0.2510	0.2891	0.3119	0.4691	0.2597	<u>0.2505</u>	0.2411
	Clark↓	2.0823	1.9757	1.8548	1.9489	1.9264	2.1576	1.7933	<u>1.7729</u>	1.8232
	Canber↓	4.7325	4.3204	4.3861	4.3853	4.7132	5.8894	4.1516	4.2061	4.1816
	KL↓	2.8146	0.8817	0.7092	0.8201	0.5804	2.4531	0.5534	<u>0.5124</u>	0.5010
	Cosine↑	0.6549	0.6372	0.7198	0.6428	0.6853	0.6004	0.7382	<u>0.7747</u>	0.7804
	Interse↑	0.4817	0.5416	0.6105	0.5719	0.5953	0.3941	0.5510	<u>0.6267</u>	0.6293
Emotion6	Cheb↓	0.3592	0.3870	0.2835	0.3156	0.3298	0.3624	0.2681	<u>0.2601</u>	0.2447
	Clark↓	1.8453	1.8952	1.5910	1.7125	1.7972	1.9517	1.6959	1.6838	1.6113
	Canber↓	4.5117	4.4308	3.5889	3.8058	3.9023	4.6164	3.4926	3.5472	3.4702
	KL↓	2.3199	1.0702	0.6383	0.7182	0.6547	0.9655	0.5102	<u>0.4916</u>	0.4636
	Cosine↑	0.6620	0.5084	0.7379	0.6918	0.6535	0.6734	0.7117	<u>0.7580</u>	0.7939
	Interse↑	0.5082	0.4115	0.6227	0.5456	0.5847	0.4942	0.6052	<u>0.6204</u>	0.6514

TABLE 5. Effects of penalty factor C.

Datasets	Criterion \ C	10 ⁻²	10 ⁻¹	1	10	10 ²	10 ³	10 ⁴	10 ⁵
Abstract	Cheb↓	0.2529	0.2402	0.2384	0.2394	0.2403	0.2394	0.2578	0.2831
	Clark↓	1.8761	1.8350	1.8284	1.8211	1.8218	1.8235	1.8284	2.0032
	Canber↓	4.2729	4.1823	4.1716	4.1786	4.1797	4.1795	4.3581	4.3415
	KL↓	0.5331	0.5045	0.5018	0.5027	0.5009	0.5113	0.5119	0.5435
	Cosine↑	0.7737	0.7782	0.7825	0.7856	0.7818	0.7800	0.7465	0.7408
	Interse↑	0.6078	0.6279	0.6354	0.6312	0.6399	0.6306	0.6325	0.6177
Emotion6	Cheb↓	0.2648	0.2452	0.2439	0.2433	0.2453	0.2421	0.2433	0.2558
	Clark↓	1.6648	1.6144	1.6127	1.6092	1.6105	1.6088	1.6114	1.6809
	Canber↓	3.6082	3.4785	3.4697	3.4710	3.4723	3.4692	3.4722	3.6063
	KL↓	0.4710	0.4662	0.4646	0.4632	0.4669	0.4682	0.4652	0.5147
	Cosine↑	0.7780	0.7902	0.7898	0.7915	0.7929	0.7955	0.7962	0.7585
	Interse↑	0.6405	0.6441	0.6435	0.6457	0.6488	0.6454	0.6568	0.6512

to positive with the same absolute value, this could induce some errors. (2) SA-BFGS achieves the worst performances, for this algorithm was especially designed for facial age estimation, it forms a special label distribution, the highest

description degree is corresponding to the chronological age, and the description degrees of both adjacent sides of the chronological age are gradually decrease, while the emotional labels don't have adjacent numerical relationship, so this

method cannot adapt emotion datasets. (3) For PT-SVM and PT-Bayes, the training set is resampled to the same size according to the weight of each example, so distribution labels are transformed into single label learning problem. In order to output the distribution of test instances, the classifier should give the probability of each label, which can be seen as description degree for the label. Bayes and SVM classifier gain poor accuracies on probability of labels compared with SVR. (4) AA- k NN is a simple but effective algorithm, it leads relative good results, this may results from the process that AA- k NN assigns the distribution to a new instance by finding the k nearest neighboring instances in the training set and then calculating the mean distribution of them. Depending on training sets, 'AA' algorithm gain adaptability on different datasets. But neural networks in AA-BP need adequate training instances, the sizes of the two datasets are too small to conduct training, so it performs worse than AA- k NN. (5) Although the two datasets are different on scale and emotion categories, our methods still show superior performance compared with the other state-of-the-art LDL algorithms, this illustrates that emotional features locate in some region rather than the holistic image, and learning the structure of outputs can improve the accuracies of distribution prediction. To give intuitive results of different methods, some instances and their predicted emotion distributions are shown in Figure 5.

In addition, we study the effects of penalty factor C . We conduct experiments on different values of C from $\{10^{-2}, 10^{-1}, 10, 10^2, 10^3, 10^4, 10^5\}$. Results are shown in Table 5. All values of C between 10^{-1} to 10^3 gave close and comparable results. The values beyond this range can't lead similar performances. In general, the promotions induced by parameter C are more weaker than the model.

VI. CONCLUSION

Image emotion analysis is an arduous work for it's subjectivity and noisy contents in images. By analysing the datasets, we find that some labels have high interdependency with others and propose a method based on Structural Learning to learn a mapping function of input features and output distributions. On the other hand, we propose to take advantage of Emotion Stimuli Map to crop the region which represents emotion most. Two representative features are studies on two distribution label datasets, and the effects of features' dimension are also validated. Extensive experiments show that discriminative features are got by our cropping method and the proposed Structural Learning method exceeds conventional LDL algorithms on most indices.

REFERENCES

- [1] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social Web," in *Proc. ACM Multimedia Conf.*, 2010, pp. 715–718.
- [2] D. Cao, R. Ji, D. Lin, and S. Li, "Visual sentiment topic model based microblog image sentiment analysis," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 8955–8968, 2016.
- [3] Z. Li, Y. Fan, W. Liu, and F. Wang, "Image sentiment prediction based on textual descriptions with adjective noun pairs," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1115–1132, 2018.
- [4] T. Rao, X. Li, X. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, Mar. 2019.
- [5] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.
- [6] X. Geng, Q. Wang, and Y. Xia, "Facial age estimation by adaptive label distribution learning," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4465–4470.
- [7] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1247–1250.
- [8] X. Geng, K. Smith-Miles, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," in *Proc. 24th AAAI Conf. Artif. Intell.*, Atlanta, GA, USA, Jul. 2010, pp. 451–456.
- [9] 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.
- [10] I. Tsochantaris, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proc. ICML*, 2004, pp. 104–112.
- [11] M. Sun, J. Yang, K. Wang, and H. Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2016, pp. 1–6.
- [12] 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.
- [13] 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.*, Sep. 2016, pp. 634–638.
- [14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.
- [15] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? Predicting the emotion stimuli map," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2016, pp. 614–618.
- [16] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 59th AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.
- [17] 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. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 308–314.
- [18] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.
- [19] K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 860–868.
- [20] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 155–161.
- [21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [22] S. Zhao, H. Yao, X. Jiang, and X. Sun, "Predicting discrete probability distribution of image emotions," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 2459–2463.
- [23] A. R. Ali, U. Shahid, M. Ali, and J. Ho, "High-level concepts for affective understanding of images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 679–687.
- [24] J. Wu, S. Zhao, V. S. Sheng, J. Zhang, C. Ye, P. Zhao, and Z. Cui, "Weak-labeled active learning with conditional label dependence for multilabel image classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1156–1169, Jun. 2017.
- [25] J. Zhang, C. Li, D. Cao, Y. Lin, S. Su, L. Dai, and S. Li, "Multi-label learning with label-specific features by resolving label correlations," *Knowl.-Based Syst.*, vol. 159, pp. 148–157, Nov. 2018.
- [26] Z. Wu, K. Markert, and S. Sharoff, "Fine-grained genre classification using structural learning algorithms," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 749–759.
- [27] A.-A. Liu, Y.-T. Su, P.-P. Jia, Z. Gao, T. Hao, and Z.-X. Yang, "Multiple/single-view human action recognition via part-induced multitask structural learning," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1194–1208, Jun. 2015.

- [28] D. Tuia, J. Muñoz-Marí, M. Kanevski, and G. Camps-Valls, "Structured output SVM for remote sensing image classification," *J. Signal Process. Syst.*, vol. 65, no. 3, pp. 301–310, 2011.
- [29] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Sep. 2005.
- [30] M. Uricár, R. Timofte, R. Rothe, J. Matas, and L. Van Gool, "Structured output SVM prediction of apparent age, gender and smile from deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun./Jul. 2016, pp. 730–738.
- [31] O. Wu, H. Zuo, W. Hu, and B. Li, "Multimodal Web aesthetics assessment based on structural SVM and multitask fusion learning," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1062–1076, Jun. 2016.
- [32] T. Rao, M. Xu, and D. Xu, "Learning multi-level deep representations for image emotion classification," 2016, *arXiv:1611.07145*. [Online]. Available: <https://arxiv.org/abs/1611.07145>
- [33] 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. 21st ACM Int. Conf. Multimedia*, 2013, pp. 223–232.
- [34] 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.
- [35] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural SVMs," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, Oct. 2009.
- [36] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM MM*, 2010, pp. 83–92.
- [37] P. Ekman, W. V. Friesen, and P. Ellsworth, "What emotion categories or dimensions can observers judge from facial behavior?" in *Emotion in the Human Face*. New York, NY, USA: Pergamon, 1972, pp. 39–55.
- [38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2014, pp. 1–14.
- [39] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
- [40] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City* vol. 1, no. 2, p. 1, 2007.
- [41] V. Campos, B. Jou, and X. Giró-I-Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vis. Comput.*, vol. 65, pp. 15–22, Sep. 2017.



YANGYU FAN received the B.S. and M.S. degrees from the Shaanxi University of Science and Technology, in 1982 and 1992, respectively, and the Ph.D. degree from Northwestern Polytechnical University, in 1999, where he is currently a Professor. His current research interests include computer vision, virtual reality, and signal processing.



HANSEN YANG received the B.S. degree in electrical engineering and its automation and the M.S. degree in measurement technology and instruments from Northwestern Polytechnical University, in 2008 and 2011, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include machine learning and image emotion analysis.



ZUHE LI received the B.S. degree in electronic information science and technology from the Zhengzhou University of Light Industry, in 2004, the M.S. degree in communication and information system from the Huazhong University of Science and Technology, in 2008, and the Ph.D. degree in information and communication engineering from Northwestern Polytechnical University, in 2017. He is currently a Lecturer with the Zhengzhou University of Light Industry. His current research interests include computer vision and machine learning.



SHU LIU received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2011 and 2017, respectively. She is currently an Assistant Professor with the School of Computer Science and Engineering, Central South University, Changsha, China. Her research interests include image processing, computer vision, machine learning, and their applications to facial attractiveness analysis.

...