Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition

Shan Li and Weihong Deng, *Member, IEEE*

***Abstract*— Facial expression is central to human experience, but most previous databases and studies are limited to posed facial behavior under controlled conditions. In this paper, we present a novel facial expression database, Real-world Affective Face Database (RAF-DB), which contains approximately 30 000 facial images with uncontrolled poses and illumination from thousands of individuals of diverse ages and races. During the crowdsourcing annotation, each image is independently labeled by approximately 40 annotators. An expectation–maximization algorithm is developed to reliably estimate the emotion labels, which reveals that real-world faces often express compound or even mixture emotions. A cross-database study between RAF-DB and CK+ database further indicates that the action units of real-world emotions are much more diverse than, or even deviate from, those of laboratory-controlled emotions. To address the recognition of multi-modal expressions in the wild, we propose a new deep locality-preserving convolutional neural network (DLP-CNN) method that aims to enhance the discriminative power of deep features by preserving the locality closeness while maximizing the inter-class scatter. Benchmark experiments on 7-class basic expressions and 11-class compound expressions, as well as additional experiments on CK+, MMI, and SFEW 2.0 databases, show that the proposed DLP-CNN outperforms the state-of-the-art handcrafted features and deep learning-based methods for expression recognition in the wild. To promote further study, we have made the RAF database, benchmarks, and descriptor encodings publicly available to the research community.**

***Index Terms*— Expression recognition, basic emotion, compound emotion, deep learning.**

I. Introduction

AUTOMATIC facial expression recognition has made significant progress in the past two decades [56], [78]. However, many developed frameworks have been employed strictly on the data collected under controlled laboratory settings with frontal faces, uniform illumination and posed expressions. On the other hand, a massive amount of images from different events and social gatherings in unconstrained environments have been captured by users from real world [15], [60]. The design of systems capable of understanding the community perception of emotional attributes and affective displays from social images is receiving increasing interest. Fortunately, the emerging deep learning techniques have advanced unconstrained expression recognition to a new state-of-the-art [31], [45]. However, to automatically infer the affective state of facial images, databases that contain large-scale valid samples and can simultaneously reflect the characteristic of real-world expressions are urgently needed.

Although Internet users [1], [6], [36], [69] provide an abundant data source for unconstrained expressions, the complexity of emotion categories annotation has hindered the collection of large annotated databases. In particular, popular AU coding [17] requires specific expertise to take months to learn and be perfected. In addition, due to cultural differences in the perception of facial emotion [18], it is difficult for psychologists to define definite prototypical AUs for each facial expression. Therefore, it is also worth to study the emotion of social images based on the judgments of a large common population by crowdsourcing [10], rather than the professional knowledge of a few experts.

Motivated by these observations, we investigate human perception and automatic recognition of unconstrained facial expressions via crowdsourcing and deep learning techniques. To this end, we have collected a large-scale, diverse, and reliably annotated facial expression database in the wild, *Real-world Affective Face Database (RAF-DB)*.[1] During annotation, 315 well-trained annotators are asked to label facial images with one of seven basic categories [16], and each image is independently annotated enough times, i.e., around 40 times in our experiment. The contributions of this paper are fourfold:

First, to enhance the readability of the label estimation, we develop an EM-based reliability estimation algorithm to evaluate the professionalism level of each labeler and then filter out the noisy labels, enabling each image to be represented reliably by a 7-dimensional emotion probability vector. By analyzing 1.2 million labels of 29,672 highly diverse facial images downloaded from the Internet, we find that real-world

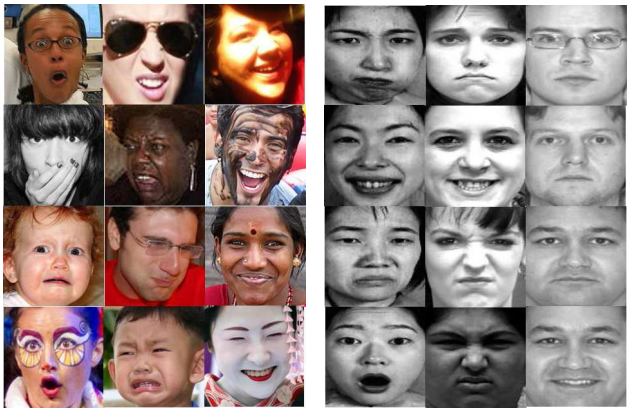
Manuscript received June 17, 2017; revised January 8, 2018 and August 20, 2018; accepted August 21, 2018. Date of publication September 3, 2018; date of current version September 19, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61573068, Grant 61471048, and Grant 61375031 and in part by the Beijing Nova Program under Grant Z161100004916088. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shiguang Shan. *(Corresponding author: Weihong Deng.)*

The authors are with the Pattern Recognition and Intelligent System Laboratory, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: ls1995@bupt.edu.cn; whdeng@bupt.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2868382

[1] http://www.whdeng.cn/RAF/model1.html



(a) (b)

Fig. 1. Examples of aligned images under real-world conditions (RAF-DB in (a)) and laboratory environments (JAFFE, CK+ and Multi-PIE in (b)). Images provided in RAF-DB are of great variability in subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions (e.g., glasses, facial hair or special gestures that hide some of the feature points), post-processing operations (e.g., various filters and special effects), etc.

affective faces are naturally categorized into two types: basic expressions with single-modal distributions and *compound emotions* with bimodal distributions. This observation supports a recent finding in the lab-controlled condition [14]. To the best of our knowledge, the real-world expression database RAF-DB is the first large-scale database to provide reliable labels of common expression perception and compound emotions in an unconstrained environment.

Second, to investigate differences between expressions captured under controlled and unconstrained conditions, we conduct a cross-database study between CK+ [46] (the most popular expression benchmark database defined by psychologists) and our RAF-DB. The expression recognition results, as well as the manual AU inspections, reveal that the AUs of real-world expressions are different from and much more diverse than those of lab-controlled expressions, as illustrated in Fig. 7. Due to these difficulties, as well as large variations in pose, illumination, and occlusion, traditional handcrafted features or shallow-learning-based features, which are well-established in lab-controlled datasets, fail to recognize facial expressions under unconstrained conditions.

Third, to improve the CNN based expression recognition, we propose a novel deep learning based framework, Deep Locality-preserving CNN (DLP-CNN). Inspired by [26], we develop a practical backpropagation algorithm that adapts the seminal idea of local neighbors from "shallow" learning to a new deep feature learning approach by creating a locality preserving loss (LP loss) which aims to pull the locally neighboring faces of the same class together. Jointly trained with the classical softmax loss which forces different classes to stay apart, locality preserving loss drives the intra-class local clusters of each class to become compact, thus highly enhancing the discriminative power of the deeply learned features. Moreover, locally neighboring faces tend to share similar emotion intensity by using DLP-CNN, which can derive discriminative deep features with smooth emotion intensity transitions. To the best of our knowledge, this is the first attempt to use such a loss function to help to supervise the learning of CNNs, thereby achieving enhanced discriminating power compared to the up-to-date approaches and setting a new state-of-the-art for expression recognition in-the-wild.

Finally, to facilitate the translation of the research from the laboratory environment to the real world, we have defined two challenging benchmark experiments on RAF-DB: 7-class basic expression classification and 11-class compound expression classification. We also conduct extensive experiments on RAF-DB and other related databases. The comparison results show that the proposed DLP-CNN outperforms handcrafted features and other state-of-the-art CNN methods. Moreover, the activation features trained on RAF-DB can be repurposed to new databases with small-sample training data, suggesting that the DLP-CNN is a powerful tool to handle the cross-culture problem on perception of emotion (POE).

This journal paper is an extended version of the conference paper [39] of CVPR 2017. The new content in this paper includes a detailed discussion of existing expression image databases, a comparative facial action coding system (FACS) analysis of CK+ and RAF-DB, a shape feature learning method for the baseline of RAF-DB and a comparative study of the proposed deep learning method on other three common databases. The remainder of this paper is structured as follows. In the next section, we briefly review related work on expression database and recognition methods. Then, we introduce the details of the construction of RAF-DB and the cross-database study between CK+ and RAF-DB in Section III. In Section IV, we introduce our new DLP-CNN approach in detail. Additionally, we include the experimental results of the baseline and DLP-CNN in Section V. Finally, we conclude and discuss future work.

II. Related Work

We first discuss related expression image datasets and then review the generic framework for facial expression analysis. Moreover, because the deep learning technique has achieved state-of-the-art performance in the field of image processing, we investigate several existing deep learning methods that have been employed for facial expression recognition.

A. Expression Image Datasets

Developments of facial expression recognition largely rely on sufficient facial expression databases. However, due to the nature of facial expression, there is a restricted number of publicly available databases providing a sufficient number of face images labeled with accurate expression information. Table I shows the summary of the existing image databases with main reference, number of samples, age range, collected environment, expression distribution, annotation method and additional information.

Several limitations among these widely used databases are common:

1. Many available databases were produced in tightly controlled environments. Subjects in them were taught to act expressions in a posed paradigm. Owing to the lack of diversity of subjects and conditions, the current

TABLE I
COMPARISON OF TEMPORAL FACIAL EXPRESSION DATABASES. VAR = VARIOUS, UNI = UNIFORM, P = POSED AND S = SPONTANEOUS

Database	Images	Age range	Envir.	Occl.	Illum.	Elicit.	Expression distribution	Annotation methods
AR [50]	4,000	N/A	Lab	Y	var	P	Smile, anger, scream and neutral	Acted by subjects
JAFFE [48]	213	N/A	Lab	N	uni	P	6 basic expressions plus neutral	Semantic ratings over 60 subjects
MMI [57]	740	19-62	Lab	Y	uni	P	6 basic expressions plus neutral	FACS coded by two coders
BU-3D [76]	2,500	18-70	Lab	N	uni	P	6 basic expressions plus neutral and four levels of intensity	Expressions acted by subjects
BU-4D [75]	606 sequences	18-45	Lab	N	uni	P	6 basic expressions plus neutral	Expressions acted by subjects
Yale [3]	165	N/A	Lab	Y	var	P	Happy, sad, sleepy, surprised, wink and normal	Expressions acted by subjects
GEMEP [2]	7,000	Avg. 29	Lab	N	uni	P	18 expressions (including rare subtle expressions)	Expressions acted by actors
CK+ [46]	593 sequences	18-50	Lab	N	uni	P&S	6 basic expressions plus contempt and neutral	FACS coded by two certified coders
Radbound [37]	8,040	N/A	Lab	N	uni	P	6 basic expressions plus contempt and neutral	Percentage of agreement on emotion categorization
Multi-PIE [23]	755,370	Avg. 27.9	Lab	Y	var	P	Smile, surprised, squint, disgust, scream and neutral	Expressions acted by subjects
BP4D [80]	328 sequences	18-29	Lab	N	uni	S	6 basic emotions plus embarrassment and pain	Self-report and rating report
FER-2013 [22]	35,887	N/A	Web	Y	var	S	6 basic expressions plus neutral	Image search API
SFEW 2.0 [12]	1,635	1-70	Movie	Y	var	P & S	6 basic expressions plus neutral	Two independent labelers per image
EmotioNet [4]	1,000,000	N/A	Web	Y	var	P & S	23 basic expressions or compound expressions	10% annotated manually and 90% annotated automatically
AffectNet [53]	450,000 (labeled)	0-50+	Web	Y	var	P & S	6 basic expressions plus neutral	One human annotator per image
RAF-DB	**29,672**	**0-70+**	**Web**	**Y**	**var**	**P & S**	**6 basic expressions plus neutral and 12 compound expressions**	**Distribution values from about 40 independent labelers per image**

recognition systems tested on these facial expression databases have reached near-perfect performance, which hinders the progress of expression recognition in the wild.

2. Images captured in real-life scenarios often present complex, compound or even ambiguous emotions rather than simple and prototypical emotions. However, the majority of the current databases include only six basic categories (surprise, fear, disgust, happiness, sadness and anger) or fewer.
3. The number of labelers in these databases is too small, which reduces the reliability and validity of the emotion labels. Additionally, emotion labels in most posed expression databases have referred to what expressions were requested rather than what was actually performed.

We then focus on discussing image databases with spontaneous expressions.

SFEW 2.0 [12] collected images from movies using key-frame extraction method and was introduced in the EmotiW 2015 Challenge. The database covers unconstrained facial expressions, varied head poses, a large age range, occlusions, varied focus and different resolutions of faces. However, it contains only 1,635 images labeled by two independent labelers.

FER-2013 [22] contains 35,887 images collected and automatically labeled by Google image search API. Cropped images are provided in 48×48 low resolution and converted to grayscale. Unfortunately, FER-2013 does not provide information about facial landmark location and the images are difficult to register well at the provided resolution and quality.

BP4D-Spontaneous [80] contains abundant images with high resolution from 41 subjects displaying a range of spontaneous expressions elicited through eight tasks. One highlight of BP4D is that it captured images using a 3D dynamic face capturing system. However, the database organization were lab-controlled, and all the subjects in this dataset are young adults.

AM-FED [51] contains 242 facial videos from the real world. Spontaneous facial expressions were captured from subjects under different recording conditions while they were watching Super Bowl commercials. The database was annotated for the presence of 14 FACS action units. However, without specific emotion labels, it is more suited for researches on AUs.

EmotioNet [4] is a large-scale database with one million facial expression images collected from the Internet. Most samples were annotated by an automatic AU detection algorithm, and the remaining 10% were manually annotated with AUs. EmotioNet contains 6 basic expressions and also 17 compound expressions; however, the emotion categories are judged based on AU label and not manually annotated.

AffectNet [53] contains more than one million images obtained from the Internet by querying different search engines using emotion related tags. A total of 450,000 images are

annotated with basic expressions by 12 labelers. Furthermore, this database contains continuous dimensional (valences and arousal) models for these images. However, each image was labeled by only one annotator due to time and budget constraints, and compound expressions are not included.

In contrast to these databases, RAF-DB simultaneously satisfies multiple requirements: sufficient data, various environments, group perception of facial expressions and data labels with minimal noise.

B. The Framework for Expression Recognition

Automatic facial expression analysis procedures can generally be divided into three main components [20]: face acquisition, facial feature extraction and facial expression classification.

In the face acquisition stage, an automatic face detector is used to locate faces in complex scenes. Feature points are then used to crop and align faces into a unified template by geometric transformations.

For facial feature extraction, previous methods can be categorized into two main groups: appearance-based methods and AU-based methods. Appearance-based methods [49] use common handcrafted feature extraction methods, such as LBP [64] and Haar [72]. AU-based methods [68] recognize expressions by detecting AUs. The most well-known AUs included in our study are the following: AU1–Inner Brow Raiser, AU2–Outer Brow Raiser, AU4–Brow Lowerer, AU5–Upper Lid Raiser, AU6–Cheek Raiser, AU7–Lid Tightener, AU9–Nose Wrinkler, AU10–Upper Lip Raiser, AU12–Lip Corner Puller, AU15–Lip Corner Depressor, AU17–Chin Raiser, AU20–Lip stretcher, AU23–Lip Tightener, AU24–Lip Pressor, AU 25–Lips part, AU26–Jaw Drop, AU 27–Mouth Stretch. Furthermore, mid-lever feature learning methods [8], [24], [44] based on manifold learning have been developed to enhance the discrimination ability of extracted low-level features.

Feature classification is performed in the final stage. The commonly used classification methods for emotion recognition include support vector machine (SVM), nearest neighbor (NN) based classifier, LDA, HMM, DBN and decision-level fusion on these classifiers [34], [78]. The extracted facial expression information is either classified as a particular facial action or a particular basic emotion [56]. Most of the studies on automatic expression recognition focus on the latter; yet, the majority of the existing systems for emotional classification is based upon Ekman's cross-cultural theory of six basic emotions [17]. Indeed, without making additional assumptions about how to determine what action units constitute an expression, there can be no exact definition for the expression category. The basic emotional expressions are therefore not universal enough to generalize expressions displayed on the human face [61].

C. Deep Learning for Expression Recognition

Recently, deep learning algorithms have been applied to visual object recognition, face verification and detection, image classification and many other problems, and have achieved state-of-the-art results. So far, few deep neural networks have been used in facial expression recognition due to the lack of sufficient training samples. In ICML 2013 competition [22], the winner [67] was based on Deep Convolutional Neural Network (DCNN) plus SVM. In EmotiW 2013 competition [11], the winner [32] combined modality specific deep neural network models. In EmotiW 2015 [12], more competitors implemented deep learning methods: transfer learning was used to solve the problem of small database in [54], a hierarchical committee of multi-column DCNNs in [33] gained the best result on SFEW 2.0 database, and LBP features combined with a DCNN structure were proposed in [38]. In [41], AU-aware Deep Networks (AUDN) was proposed to learn features with the interpretation of facial AUs. In [42], 3D Convolution-al Neural Networks (3DCNN-DAP) with deformable action parts constraints were adopted to localize the action parts and encode them effectively. In DTAGN [30], two different models were combined to extract temporal appearance and geometry features simultaneously. In [52], a DCNN with inception layers was proposed to achieve comparable results. In [79], a DNN-driven feature learning method was proposed to address multi-view facial expression recognition.

TABLE II
KEYWORDS USED TO COLLECT THE IMAGES FOR RAF-DB

Joy	Sadness	Anger	Fear	Surprise	Disgust	Neutral
smile	sad	angry	scared	surprised	disgust	straight
laugh	annoyed	anger	frightened	omg	disgusted	portrait
giggle	cry	pissed-off	fear	shocked		potrait
big smile	crying	rage	horrified	astonished		
	depressed		fearful	amazed		
	heartbroken		afraid	surprised		
	disappointed		terrified			

III. REAL-WORLD EXPRESSION DATABASE: RAF-DB

A. Creating RAF-DB

1) Data collection: At the very beginning, images' URLs collected from Flickr were fed into an automatic open source downloader to download images in batches. Considering that the results returned by Flickr's image search API were in well-structured XML format, from which the URLs can be easily parsed, we then used a set of keywords to pick out the images that were related to the six basic emotions plus the neutral emotion. At last, a total of 29,672 real-world facial images are presented in our database. Some of the emotion-related keywords used are listed in Table II. Figure 2 shows the pipeline of data collection.

2) Database Annotation: Annotating nearly 30,000 images is an extremely difficult and time-consuming task. Considering the compounded property of real-world expressions, multiple views of images' expression states should be collected from different labelers. Therefore, we employed 315 annotators (students and staff from universities) who were instructed with a one-hour tutorial of psychological knowledge on emotion for an online facial expression annotation assignment, during

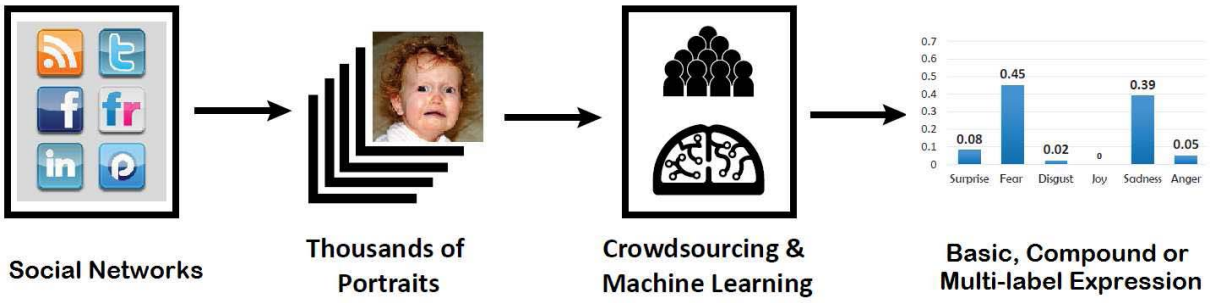
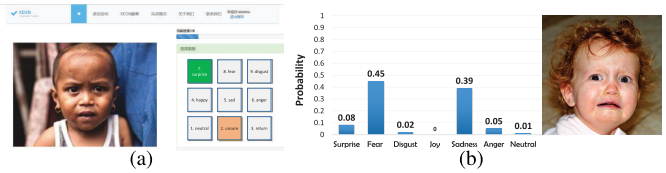
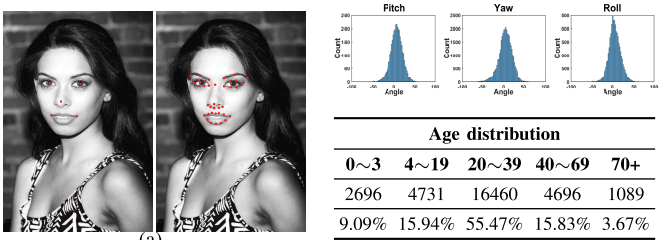


Fig. 2. Overview of the construction and annotation of RAF-DB. Initially, images collected from Flickr were fed into an automatic downloader to download the images in batches. Then, a large number of real-world facial images were picked out using emotion-related keywords. To guarantee the reliability of the labeling results, we have invited sufficient well-trained labelers to independently annotate each image about 40 times.



(a) (b)

Fig. 3. Database annotations. (a): The web-based framework used for the annotation. (b): A Sadly Fearful sample from RAF-DB with its 7-dimensional expression distribution.



(a)

Age distribution				
0∼3	4∼19	20∼39	40∼69	70+
2696	4731	16460	4696	1089
9.09%	15.94%	55.47%	15.83%	3.67%

(b)

Fig. 4. (a) Sample face with five accurate landmarks manually corrected by our experimenters and 37 landmarks automatically annotated using Face++ API; (b) Age and pose distributions of the images in RAF-DB.

which they were asked to classify images into the most apparent one from seven classes. We developed a website to make it easy for our annotators to contribute, which shows each image with exclusive attribute options. Images were randomly and equally assigned to each labeler, ensuring that there was no direct correlation among the images labeled by one person. And each image was ensured to be labeled by about 40 independent labelers. After that, a multi-label annotation result was obtained for each image, i.e., a seven-dimensional vector where each dimension indicates the votes for the relevant emotion. The UI of the database annotation application is shown in Figure 3(a), and an example of a typical annotation result is shown in Figure 3(b).

3) Metadata: The data are provided with precise locations and the size of the face region, as well as five manually located landmark points (the centers of two eyes, the tip of the nose and the two corners of the mouth) on the face. The 5 facial landmarks were accurately located by labelers who were asked to guess the locations of occluded landmarks. Specifically, the rough locations and points were first detected automatically using the Viola-Jones face detector [70] and SDM [74] based methods. Then, imprecise or missed detections and localizations were corrected by human labelers. Besides, an automatic landmark annotation mode without manual label is included: 37 landmarks were picked out from the annotation results provided by Face++ API [28]. Figure 4(a) shows sample faces with five precise landmarks and 37 landmarks. We also manually annotated the basic attributes (gender, age (5 ranges) and race) of all RAF faces. In summary, subjects in our database range in age from 0 to 70 years old. They are 52% female, 43% male, and 5% unclear. For racial distribution, there are 77% Caucasian, 8% African American, and 15% Asian. The pose of each image, including pitch, yaw and roll parameters, is computed from the manually labeled locations of the five facial landmarks. Figure 4(b) shows the age (images with unclear gender are of infants.) and pose distributions in RAF-DB.

4) Reliability Estimation: Due to the subjectivity and varied expertise of the labelers and the wide range of image difficulty, there was some disagreement among annotators. To get rid of noisy labels, motivated by [73], an Expectation Maximization (EM) framework was used to assess each labeler's reliability.

Let $\mathcal{D} = \{(x_j, y_j, t_j^1, t_j^2, \ldots, t_j^R)\}_{j=1}^n$denote a set of n labeled inputs, where y_j is the gold standard label (hidden variable) for the jth sample x_j, and $t_j^i \in \{1, 2, 3, 4, 5, 6, 7\}$ is the corresponding label given by the ith annotator. The correct probability of t_j^i is formulated as a sigmoid function $p(t_j^i = y_j|\alpha_i, \beta_j) = (1 + \exp(-\alpha_i\beta_j))^{-1}$, where $1/\beta_j$ is the difficulty of the jth image and α_i is the reliability of the ith annotator.

Our goal is to optimize the log-likelihood of the given labels:

$$\begin{aligned}\max_{\beta>0} l(\alpha, \beta) &= \sum_j \ln p(t|\alpha, \beta) = \sum_j \ln \sum_y p(t, y|\alpha, \beta) \\ &= \sum_j \ln \sum_y Q_j(y) \frac{p(t, y|\alpha, \beta)}{Q_j(y)} \\ &\geq \sum_j \sum_y Q_j(y) \ln \frac{p(t, y|\alpha, \beta)}{Q_j(y)},\end{aligned}$$

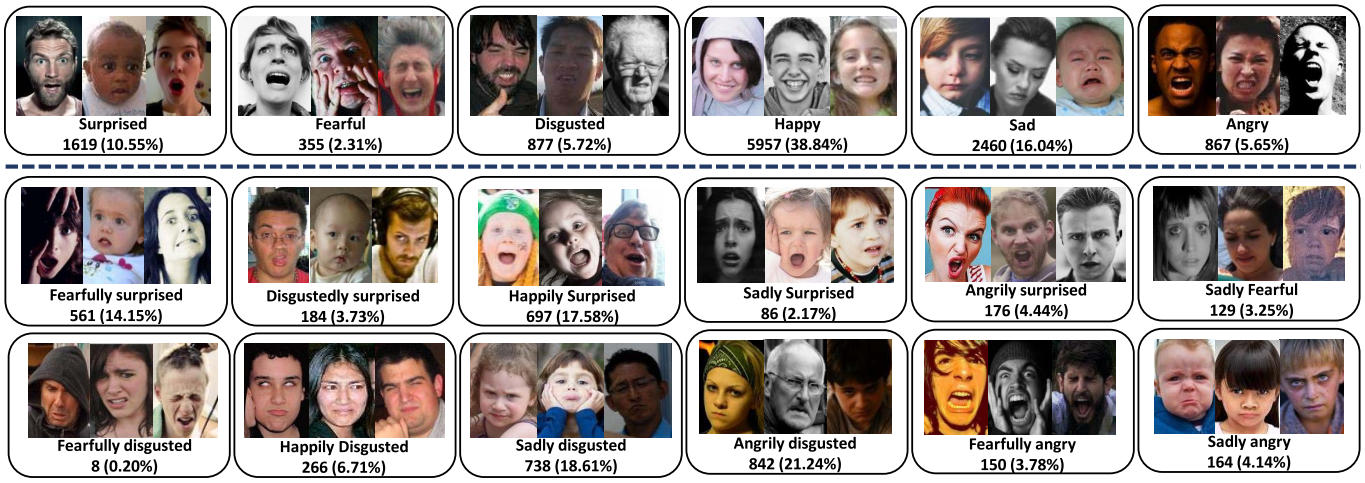


Fig. 5. Examples of six-class basic emotions and twelve-class compound emotions from RAF-DB. The detailed data proportion and class distribution of RAF-DB are attached to each expression class.

Algorithm 1 Label Reliability Estimation Algorithm

Input: Training set $\mathcal{D} = \{(x_j, t_j^1, t_j^2, ..., t_j^R)\}_{j=1}^n$
Output: Each annotator's reliability α_i^*
Initialize:
$\forall j = 1, ..., n$, initialize the true label y_j using majority voting

$$\beta_j := -\sum_{i=1}^{R} p(t_j^i) \ln p(t_j^i),\ \alpha_i := 1,$$

The initial value of β_j is image j's entropy. The higher the entropy, the more uncertain the image.
Repeat:
E-step:

$$Q_j(y_j) := \prod_i p(y_j | t_j, \alpha_i, \beta_j)$$

M-step:

$$\alpha_i := \arg\max_{\alpha_i} \sum_j \sum_{y_j} Q_j(y_j) \ln \frac{p(t_j, y_j | \alpha_i, \beta_j)}{Q_j(y_j)}$$

We also optimize β_j along with α_i during M-step. However, the goal is to get each labeler's reliability, so we didn't include it in this step. For optimization, we take a derivative with respect to β_j and α_i respectively.
Until convergence

where $Q_j(y)$ is a certain distribution of hidden variable y,

$$Q_j(y_j) = \frac{p(t_j, y_j | \alpha, \beta)}{\sum_y p(t_j, y_j | \alpha, \beta)} = \frac{p(t_j, y_j | \alpha, \beta)}{p(t_j | \alpha, \beta)} = p(y_j | t_j, \alpha, \beta).$$

After revision, 285 annotators' labels have been remained and Cronbach's Alpha score of all labels is 0.966. Algorithm 1 summarizes the learning process of label reliability estimation. In contrast to the Gaussian prior initialization in [73], we further introduced the prior knowledge of annotation for faster convergence.

5) Subset Partitions: Let $G_j = \{g_1, g_2, \ldots, g_7\}$ denote the 7-dimensional ground truth of the jth image, where $g_k = \sum_{i=1}^{R} \alpha_i 1_{t_j^i = k}$ (α_i means the ith annotators' reliability. 1_A is an indicator function that evaluates to "1" if the Boolean expression A is true and "0" otherwise.), and label $k \in \{1, 2, 3, 4, 5, 6, 7\}$ refers to surprise, fear, disgust, happiness, sadness, anger and neutral, respectively. We then divide RAF-DB into different subsets according to the 7-dimensional ground truth. For the Single-label Subset, we first calculate the mean distribution value $g_{mean} = \sum_{k=1}^{7} g_k/7$ for each image, then select label k w.r.t. $g_k > g_{mean}$ as the valid label. Images with a single valid label are classified into Single-label Subset. For Two-tab Subset, the partition rule is similar. The only difference is that we removed images with neutral labels before the partition step. Figure 5 exhibits specific samples and the concrete proportion of 6-class basic emotions and 12-class compound emotions.

B. CK+ and RAF Cross-Database Study

We then conducted a CK+ [46] and RAF cross-database study to explore the specific differences between expressions of real-world affective faces and lab-controlled posed faces guided by psychologists. Here, "cross-database" means we use the images from one database for training and images from the other for testing. By this study, we aim to identify the real challenges of real-world affective face analysis. To eliminate the bias caused by different training sizes, the single-tab subset of RAF-DB has been sub-sampled to balance the size of these two databases.

To ensure the generalizability of the classifiers, we applied SVM for classification and implemented the HOG descriptor [9] for representation. Specifically, facial images were first aligned by an affine transformation defined by the centers of the two eyes and the center of the two corners of the mouth and then normalized to the size of 100×100. Then, the HOG features were extracted for each aligned face image. Finally, SVM with radial basis function (RBF) kernel implemented by LibSVM [7] was applied for classification. The parameters were optimized via grid search.

We then performed a cross-database experiment based on the six-class expression. Multiclass SVM (mSVM) and

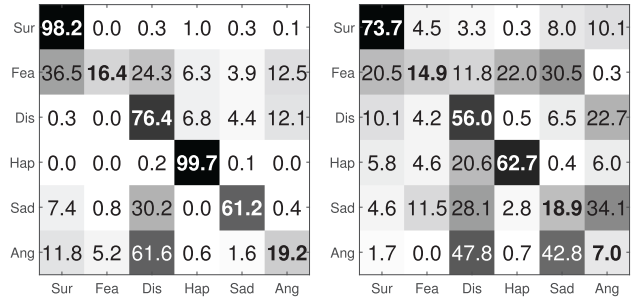


Fig. 6. Confusion matrixes for cross-database experiments using HOG features. The true labels (training data) are on the vertical axis, and the predicted labels (test data) are on the horizontal axis. (a) RAF → CK+. (b) CK+ → RAF.

confusion matrix were used as the classification method and assessment criteria, respectively. Figure 6 shows the results of this experiment, where Matrix (a) refers to training on RAF-DB and testing on CK+ and Matrix (b) refers to training on CK+ and testing on RAF-DB.

Analyzing the diagonals of these two matrixes, we can see that surprise, happiness and disgust are the emotions with the highest recognition rates in both cases. This result is in line with many single-database tests based on CK+, such as [46], [58], and [64]. The average of the diagonals indicates that Matrix (a) was detected with 62% accuracy while Matrix (b) was detected with only 39% accuracy, which indicates that data collected from the real world are more varied and effective than lab-controlled data. This is particularly evident in the expression of sadness, happiness and surprise. Furthermore, anger and disgust are often confused with each other in both cases, which conforms to the survey in [5].

To explain the phenomena above, more detailed research must be conducted to identify the specific differences in each expression between these two databases. Therefore, a facial action coding system (FACS) analysis was employed on the experimental data from RAF-DB. FACS was first presented in [17], where the changes in facial behaviors were described by a set of action units (AUs). To ensure the reliability, two FACS coders were employed to label AUs for the 309 images randomly chosen from RAF-DB. During annotation, the magnified original color facial images were displayed on the screen. The inter-observer agreement quantified by coefficient kappa was 0.83, and the two coders discussed with each other to arbitrate the disagreements. We then quantitatively analyzed the AU presence for different emotions in CK+ and RAF. Some examples from CK+ and RAF are shown in Figure 7. Additionally, the AU occurrence probabilities for each expression from the subset of RAF-DB are shown in Table III.

TABLE III
AU OCCURRENCE PROBABILITIES FOR EACH EXPRESSION IN RAF-DB

(%)	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU15	AU 17	AU20	AU25	AU 26	AU27
Sur	96	96		90									95	45^*	
Fea	82	39	72	81		42						35^*	83		53^*
Dis			49				28^*	92^*			62		63^*		
Hap					98				95				97	24	13
Sad	93		78							17^*	42		46^*		
Ang			97	79^*		69		75			56		72^*	58^*	

Missing data indicates the probability is less than 10%.

An asterisk(*) indicates the AU's probability is quite different from CK+'s (at least 40% disparity).

IV. DEEP LOCALITY-PRESERVING FEATURE LEARNING

In addition to the difficulties such as variable lighting, poses and occlusions, real-world affective faces pose at least two challenges that require new algorithms to address. First, as indicated by our cross-database study, real-world expressions may associate with various AU combinations that require classification algorithms to model the multi-modality distribution of each emotion in the feature space. Second, as suggested by our crowdsourcing results, a large proportion of real-world affective faces express compound or even multiple emotions. Therefore, traditional hand-engineered representations that perform well on laboratory-controlled databases are not suitable for expression recognition tasks in the wild.

Recently, DCNN has been proved to outperform handcrafted features on large-scale visual recognition tasks. Conventional DCNN commonly uses the softmax loss layer to supervise the training process. By denoting the i-th input feature x_i with the label y_i, the softmax loss can be written as

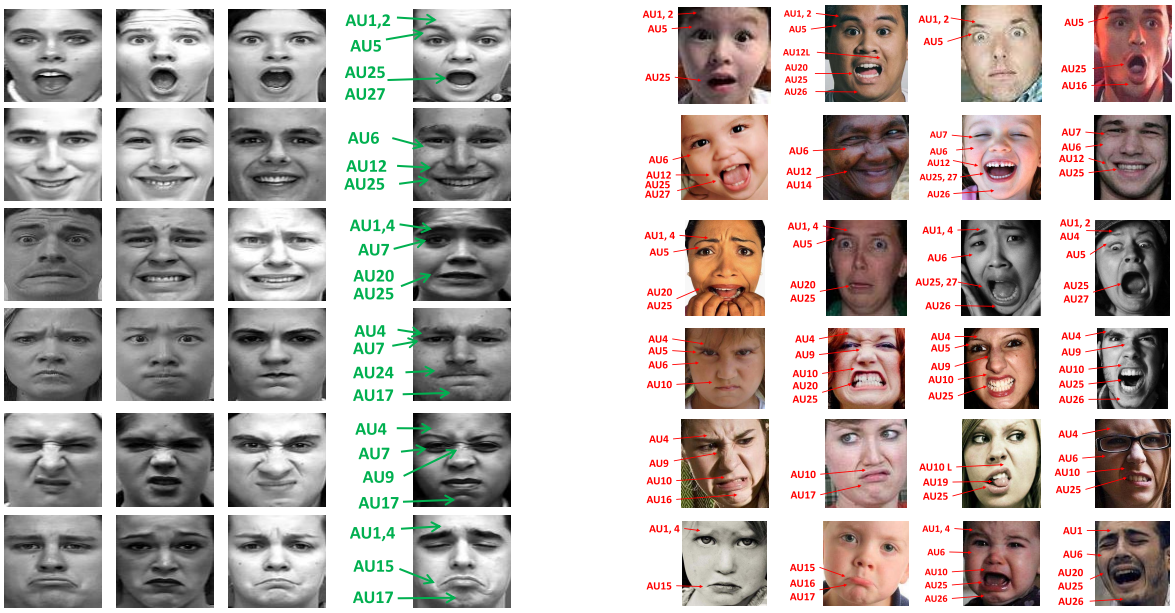
$$L_s = -\frac{1}{n}\sum_{i=1}^{n} log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}},\right) \tag{1}$$

where f_j denotes the j-th element ($j = 1 \ldots C$, C is the number of classes) of the vector of the class scores f, and n is the number of training data. The softmax layer merely helps to keep the deeply learned features of different classes (expressions) separable. Unfortunately, as shown in Figure 1, facial expressions in the real world show significant intra-class differences in occlusion, illumination, resolution and head position. Moreover, individual differences can also lead to major differences in the same expression category, for example, laugh vs. smile.

To address these difficulties, we propose a novel DLP-CNN to address the ambiguity and multi-modality of real-world facial expressions. In DLP-CNN, we add a new supervised layer to the fundamental architecture shown in Table IV, namely, locality preserving loss (LP loss), to improve the discrimination ability of the deep features. The basic idea is to preserve the locality of each sample x_i and to make the local neighborhoods within each class as compact as possible. To formulate our goal:

$$\min \sum_{i,j} S_{ij} \left\| x_i - x_j \right\|_2^2, \tag{2}$$

where matrix S is a similarity matrix. The deep feature $x \in \mathbb{R}^d$ denotes deep convolutional activation features (DeCaf) [13] taken from the final hidden layer, i.e., just before the softmax layer that produce the class predictions. A possible way to



(a) (b)

Fig. 7. Comparison of six basic emotions from CK+ and RAF. Facial expressions from top to bottom are Surprise, Happiness, Fear, Anger, Disgust and Sadness. It is evident that the expression AUs in RAF are more diverse than those in CK+. (a) CK+. (b) RAF-DB.

TABLE IV
THE CONFIGURATION PARAMETERS IN THE FUNDAMENTAL ARCHITECTURE (BASEDCNN)

Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Type	Conv	ReLu	MPool	Conv	ReLu	MPool	Conv	ReLu	Conv	ReLu	MPool	Conv	ReLu	Conv	ReLu	FC	ReLu	FC
Kernel	3	-	2	3	-	2	3	-	3	-	2	3	-	3	-		-	
Output	64	-	-	96	-	-	128	-	128	-	-	256	-	256	-	2,000	-	7
Stride	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1		1	
Pad	1	0	0	1	0	0	1	0	1	0	0	1	0	1	0		0	

define S is as follows.

$$S_{ij} = \begin{cases} 1, & x_j \text{ is among the } k\text{-nearest neighbors of } x_i \\ & \text{or } x_i \text{ is among the } k\text{-nearest neighbors of } x_j \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where x_i and x_j belong to the same class of expression and k defines the size of the local neighborhood.

This formulation effectively characterizes the intra-class local scatter. Note that x_i should be updated as the iterative optimization of the CNN. To compute the summation of the pairwise distances, we need to consider the entire training set in each iteration, which is inefficient to implement. To address this difficulty, we do the approximation by searching the k-nearest neighbors for each sample x_i, and the LP loss function of x_i is defined as follows:

$$L_{lp} = \frac{1}{2n} \sum_{i=1}^{n} \left\| x_i - \frac{1}{k} \sum_{x \in N_k\{x_i\}} x \right\|_2^2, \tag{4}$$

where $N_k\{x_i\}$ denotes the ensemble of the k-nearest neighbors of sample x_i with the same class.

The gradient of L_{lp} with respect to x_i is computed as:

$$\frac{\partial L_{lp}}{\partial x_i} = \frac{1}{n} \left(x_i - \frac{1}{k} \sum_{x \in N_k\{x_i\}} x \right). \tag{5}$$

In this manner, we can perform the update based on mini-batch. Note that the recently proposed center loss [71] can be considered to be a special case of the LP loss if $k = n_c - 1$ (n_c is the number of training samples in class c to which x_i belong). While center loss simply pulls the samples to a single centroid, the proposed LP loss is more flexible, especially when the class conditional distribution is multi-modal.

We then adopt the joint supervision of softmax loss, which characterizes the global scatter, and the LP loss, which characterizes the local scatters within class, to train the CNNs for discriminative feature learning. The objective function is formulated as follows: $L = L_s + \lambda L_{lp}$, where L_s denotes the softmax loss and L_{lp} denotes the LP loss. The hyperparameter λ is used to balance the two loss functions. Algorithm 2 summarizes the learning process in the DLP-CNN. Intuitively, *the softmax loss forces the deep features of different classes to remain apart and the LP loss efficiently pulls the neighboring deep features of the same class together.* With the joint supervision, both the inter-class feature differences and the intra-class

Algorithm 2 Optimization Algorithm of DLP-CNN

Input: Training data $\{x_i, y_i\}_{i=1}^n$,
n is the size of the mini-batch
Output: Network layer parameters W
Initialize: $t \leftarrow 0$
Network learning rate μ, hyperparameter λ, network layer parameters W, softmax loss parameters θ, neighboring nodes k.
Repeat:
1: $t \leftarrow t + 1$
2: Compute the center of the k-nearest neighbor for x_i:
$C_i^t = \frac{1}{k} \sum_{j=1}^n x_j^t S_{ij}^t$
3: Update the softmax loss parameters:
$\theta^{t+1} = \theta^t - \mu^t \frac{\partial L_s^t}{\partial \theta^t}$
4: Update the backpropagation error:
$\frac{\partial L^t}{\partial x_i^t} = \frac{\partial L_s^t}{\partial x_i^t} + \lambda \frac{\partial L_{lp}^t}{\partial x_i^t}$
5: Compute the network layer parameters:
$W^{t+1} = W^t - \mu^t \frac{\partial L^t}{\partial W^t} = W^t - \mu^t \sum_{i=1}^n \frac{\partial L^t}{\partial x_i^t} \frac{\partial x_i^t}{\partial W^t}$
Until convergence

feature correlations are enlarged. Hence, the discriminative power of the deeply learned features can be highly enhanced.

V. Experimental results on baseline and DLP-CNN

We first conducted baseline experiments on RAF-DB. Then, the proposed deep learning method, Deep Locality-preserving CNN, was employed to solve the difficulties of facial expression recognition in real world.

A. Baseline Experiment

In this section, to evaluate the tasks with real-world challenges, we performed two benchmark experiments on RAF-DB and presented affiliated baseline algorithms and performances. While our main purpose is to analyze the results of the aforementioned techniques on RAF-DB, we also conduct experiments on two small and popular datasets, CK+ and JAFFE [48].

For facial representations, we employ two types of information encoded in the feature space: shape and appearance. Experiments on human subjects demonstrate that shape representations play a role in the recognition of the emotion class from face images [27], [62]. Before computing our feature space, all images are aligned and downsized to 100*100 pixels within the given precise five landmarks.

The 37 fiducial points are used to determine the dimensions of our shape feature. More formally, given two fiducial points, z_i and z_j, where $i \neq j$, i and $j = \{1, \ldots, 37\}$, $z_i = (z_{i1}, z_{i2})^T$, z_{i1} and z_{i2} are the horizontal and vertical components of the fiducial point, respectively, and their relative positions are $d_{ijk} = z_{ik} - z_{jk}$, $k = 1, 2$. With these 37 facial landmarks, the feature vector $\mathbf{f}$ has $2 \cdot (37 \cdot 36)/2 = 1,332$ dimensions defining the shape of the face. Before passing the features into the classifier, we normalize the feature vector $\mathbf{f}$ to be $\tilde{\mathbf{f}}$ as follows:

$$\tilde{f}_i = \frac{1}{2}\left(\frac{f_i - \mu_i}{2\sigma_i} + 1\right), \tag{6}$$

where $i = \{1, \ldots, 1332\}$, and μ_i and σ_i are the mean and standard deviation of the ith feature across the training data, respectively. Then, we truncate the out-of-range elements to either 0 or 1.

We also employ three widely used low-level appearance features: LBP [55], HOG [9] and Gabor [40] representations. For LBP, we select the 59-bin $LBP_{8,2}^{u2}$ operator and divide the 100*100 pixel images into 100 regions with a 10*10 grid size, which was empirically found to achieve relatively good performance for expression classification. For HOG, we first divide the images into 10*10 pixel blocks of four 5*5 pixel cells with no overlapping. By setting 10 bins for each histogram, we obtain a 4,000-dimensional feature vector per aligned image. For the Gabor wavelet, we use a bank of 40 Gabor filters at five spatial scales and eight orientations. The downsampled image size is set to 10*10, yielding 4,000-dimensional features.

For the classification task we use linear SVMs with a one-against-one strategy to decompose the multi-class classification problem into multiple binary-class classifications by voting. Given a training set $\{(x_i, y_i), i = 1, \ldots, n\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, any test sample x can be classified using:

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n max(1 - y_i w^T x_i, 0). \tag{7}$$

To objectively measure the performance for the followers entries, we split the dataset into a training set and a test set, where the training set is five times larger than the test set, and the expressions in both sets have a near-identical distribution. Because real-world expressions have an imbalanced distribution, the accuracy metric, which is employed as the evaluation criterion in most datasets, is not used in RAF as it is especially sensitive to bias and is not effective for imbalanced data [21]. Instead, we use the mean diagonal value of the confusion matrix as the metric. During the parameter optimization process, we also optimized the mean diagonal value of the confusion matrix rather than the accuracy directly provided by the SVMs.

1) Basic Emotions: In this experiment, seven basic emotion classes were detected using all 15,339 images from the single-label subset. The best classification accuracy (output by SVM) was 66.82% for shape features, 72.71% for LBP, 74.35% for HOG, and 77.28% for Gabor. The accuracy results decreased to 50.52%, 55.98%, 58.45% and 65.12% when using the mean diagonal value of the confusion matrix as the metric. To assess the reliability of the basic emotion labels, we also assigned a uniform random label, which we call a naive emotion detector, to each sample. The best result for the naive classifier was 16.07% when using the Gabor feature, which is much lower than the former value.

For comparison, we employed the same methods on CK+ with 5-fold cross-validation and JAFFE with a leave-one-subject-out strategy. The results shown in Table V confirm that real-world expressions are more difficult to recognize and the current common methods that perform well on the existing databases cannot solve the expression recognition problem in challenging real-world conditions.

TABLE V
BASIC EXPRESSION CLASS PERFORMANCE COMPARISON OF CK+, JAFFE AND RAF ALONG WITH COMPOUND EXPRESSION PERFORMANCE OF RAF BASED ON THE LBP, HOG AND GABOR DESCRIPTORS, AND SVM AND LDA+KNN CLASSIFICATION. THE METRIC IS THE MEAN DIAGONAL VALUE OF THE CONFUSION MATRIX

		basic			compound
		CK+	JAFFE	**RAF**	**RAF**
mSVM	shape	–	–	**50.52**	**28.84**
	LBP	88.92	78.81	**55.98**	**28.84**
	HOG	90.50	84.76	**58.45**	**33.65**
	Gabor	91.98	88.95	**65.12**	**35.76**
LDA	shape	–	–	**42.87**	**20.44**
	LBP	85.84	77.74	**50.97**	**22.89**
	HOG	91.77	80.12	**51.36**	**24.01**
	Gabor	92.33	83.45	**56.93**	**23.81**

To evaluate the effectiveness of different classifiers, we have also trained LDA with nearest neighbor (NN) classification. We found that LDA+NN was inferior to mSVM when training on RAF, an extremely large database. Nevertheless, LDA+NN performed better when training on small-scale datasets (CK+ and JAFFE), even outperforming mSVM in some cases. The concrete results are given in Table V.

2) Compound Emotions: As suggested by our crowdsourcing results, a large proportion of real-world affective faces express compound emotions. The performance evaluation on the single-emotional data set may not be sufficiently comprehensive for some real-world applications. Therefore, we conducted additional baseline experiments on compound emotions.

For compound emotion classification, we removed the fearfully disgusted emotion due to an insufficient number of samples, leaving 11 classes of compound emotions, 3,954 in total. The best classification accuracy (output by SVM) was 45.96% for shape features, 45.51% for LBP, 51.89% for HOG, and 53.54% for Gabor. The accuracy decreased to 28.84%, 28.84%, 33.65% and 35.76% when using the mean diagonal value of the confusion matrix as the metric. Again, to demonstrate the reliability of the compound emotion labels, we computed the baseline for the naive emotion detector, which decreased to 5.79% when using the Gabor features.

As expected, the overall performance decreased substantially when more expressions were included in the classification. The significantly worse results compared to those of the basic emotion classification indicate that compound emotions are more difficult to detect and that new methods should be invented to solve this problem. In addition to the multi-modality, the lack of training samples for compound expressions from the real world is another major technical challenge.

B. Deep Learning Experiment

Nowadays, deep learning has been applied to large-scale visual recognition tasks and has performed exceedingly well with large amounts of training data. However, fully supervised deep models are easy to be overfitting on facial expression recognition tasks due to the insufficient training samples for model training. Therefore, most deep learning frameworks employed for facial expression recognition [38], [54], [59] are based on pre-trained models. These pre-trained models, such as the VGG network [66] and AlexNet [35], were initially designed for face recognition, which are short of discrimination ability of expression characteristic. Therefore, in this paper, we directly trained our deep learning system on the sufficiently large self-collected RAF-DB from scratch without using other databases.

All our models were trained based on the open source deep learning framework, Caffe [29]. The already aligned grayscale images were first normalized by dividing all the pixel values by 255. We then considered a network taking a fixed-size input (90*90) cropped from the images for data augmentation.

To compare different methods fairly, we adopted uniform training methods and used uniform fundamental network architectures. The learning rate was initially set to 0.01 and was decreased by a factor of 10 at 5k and 18k iterations, and we stopped training at 20k iterations. Moreover, we chose stochastic gradient descent (SGD) for optimization and used mini-batch with 64 samples. The momentum coefficient was set to 0.9.

The model was regularized using weight decay. We set the weight decay coefficient of the convolutional layer and first fully connected layer to 0.0005 and that of the second fully connected layer to 0.0025. MSRA [25] was used to initialize the weight parameter of the convolutional layer and fully connected layer, while the bias parameter was set to 0 at the beginning of training. All our models were trained on an NVIDIA Tesla K40 GPU, and approximately 3 hours was required to train a model.

When conducting the experiments, we followed the same dataset partition standards, image processing methods and classification methods as those of the baseline system. Related research [13] proved that well-trained deep convolutional network can work as a feature extraction tool with generalizability for the classification task. Following up this idea, we first trained each DCNN for the basic emotion recognition task (that is, we used the basic emotion training set as the training samples and the basic emotion test set as the validation samples) and then directly used the already trained DCNN models to extract deep features for both basic and compound expressions. The 2,000-dimensional deep features learned from the raw data were extracted from the penultimate fully connected layer of the DCNNs and were then classified by SVM.

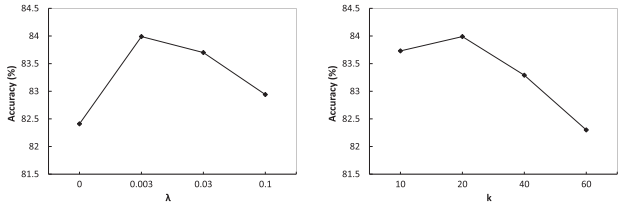
To investigate the efficiency of different values of λ and k used in the DLP-CNN model, we conducted two experiments on the basic expression recognition task. The accuracies predicted directly by DLP-CNN for the basic expression recognition are shown in Figure 8. In the first experiment (left), we fixed k to 20 and varied λ from 0 to 0.1 to train different models. As the results show, the accuracies are sensitive to the choice of λ and $\lambda = 0$ is the case of using the softmax loss, which leads to relatively poor performance of the deeply

TABLE VI

EXPRESSION RECOGNITION PERFORMANCE OF DIFFERENT DCNNS ON RAF. THE METRIC IS THE MEAN DIAGONAL VALUE OF THE CONFUSION MATRIX

		basic									compound	
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral	**Average**†	Accuracy⋆	**Average**†	Accuracy⋆
mSVM	VGG [66]	68.52	27.50	35.13	85.32	64.85	66.32	59.88	**58.22**	70.53	**31.63**	49.62
	AlexNet [35]	58.64	21.87	39.19	86.16	60.88	62.31	60.15	**55.60**	68.90	**28.22**	45.45
	baseDCNN	70.99	52.50	50.00	92.91	77.82	79.64	83.09	**72.42**	82.86	**40.17**	56.69
	center loss [71]	68.52	53.13	54.05	93.08	78.45	79.63	83.24	**72.87**	83.68	**39.97**	55.81
	DLP-CNN	71.60	52.15	62.16	92.83	80.13	81.16	80.29	**74.20**	84.13	**44.55**	57.95
LDA	VGG [66]	66.05	25.00	37.84	73.08	51.46	53.49	47.21	**50.59**	58.15	**16.27**	27.55
	AlexNet [35]	43.83	27.50	37.84	75.78	39.33	61.70	48.53	**47.79**	57.43	**15.56**	26.41
	baseDCNN	66.05	47.50	51.35	89.45	74.27	76.90	77.50	**69.00**	78.75	**28.23**	38.15
	center loss [71]	64.81	49.38	54.05	92.41	74.90	76.29	77.21	**69.86**	78.91	**27.33**	37.46
	DLP-CNN	77.51	55.41	52.50	90.21	73.64	74.07	73.53	**70.98**	79.95	**32.29**	42.93

† The mean diagonal value of the confusion matrix.
⋆ The accuracy directly output by SVM.



(a) (b)

Fig. 8. Basic expression recognition performance on RAF-DB for different values of λ (left) and k (right). (a) DLP-CNN models with different λ and fixed $k = 20$. (b) DLP-CNN models with different k and fixed $\lambda = 0.003$.

TABLE VII

CONFUSION MATRIX FOR THE SEVEN BASIC EMOTION CATEGORIES WHEN USING DLP-CNN

(%)	Sur	Fea	Dis	Hap	Sad	Ang	Neu
Sur	**81.16**	1.52	1.22	5.17	2.74	1.52	6.69
Fea	9.46	**62.16**	2.70	8.11	6.76	4.05	6.76
Dis	1.88	1.25	**51.25**	9.38	10.63	8.75	16.88
Hap	0.84	0.34	0.42	**92.83**	1.18	0.59	3.80
Sad	0.42	0.84	2.30	6.28	**80.13**	0.84	9.21
Ang	4.32	3.09	6.79	6.79	3.70	**71.60**	3.70
Neu	2.06	0.00	2.65	5.44	8.97	0.59	**80.29**

learned features. In the second experiment (right), we fixed $\lambda = 0.003$ and varied k from 10 to 60 to train different models, and we achieved the best performance when k was set to 20.

Concrete classification results of the basic and compound expression in RAF-DB when using DLP-CNN features are shown in Table VII and Table VIII. Figure 9(b) shows the resulting 2-dimensional deep features learned from our DLP-CNN model, where we attach example face images with various intensity in different expression classes. Although the RAF images include various identities, poses, and lighting, the face images are mapped into a two-dimensional space with separable expression clusters and continuous change in expression intensity. This is because while trying to preserve the local structure of the deep features, DLP-CNN implicitly emphasizes the natural clusters in the data and preserves the smooth change within clusters. With its neighborhood-preserving character, the deep features are able to capture the intrinsic expression manifold structure to a large extent.

TABLE VIII

CONFUSION MATRIX FOR COMPOUND EMOTION CATEGORIES WHEN USING DLP-CNN. A, HAPPILY SURPRISED; B, HAPPILY DISGUSTED; C, SADLY FEARFUL; D, SADLY ANGRY; E, SADLY SURPRISED; F, SADLY DISGUSTED; G, FEARFULLY ANGRY; H, FEARFULLY SURPRISED; I, ANGRILY SURPRISED; J, ANGRILY DISGUSTED; K, DISGUSTEDLY SURPRISED

	a	b	c	d	e	f	g	h	i	j	k
a	**79.3**	4.4	0.7	0.0	0.0	0.7	0.0	10.4	1.5	3.0	0.0
b	12.8	**38.3**	0.0	2.1	0.0	10.6	0.0	2.1	2.1	31.9	0.0
c	4.5	9.1	**31.8**	4.5	0.0	18.2	0.0	22.7	4.5	4.5	0.0
d	0.0	6.1	6.1	**30.3**	0.0	24.2	12.1	0.0	0.0	21.2	0.0
e	0.0	0.0	16.7	0.0	**22.2**	27.8	0.0	16.7	5.6	11.1	0.0
f	5.0	5.0	2.8	0.7	0.0	**64.5**	0.0	1.4	0.0	18.4	2.1
g	6.1	0.0	0.0	3.0	0.0	0.0	**63.6**	15.2	3.0	9.1	0.0
h	22.4	0.9	3.4	0.0	0.9	3.4	1.7	**62.1**	1.7	2.6	0.9
i	21.1	0.0	0.0	5.3	0.0	5.3	5.3	15.8	**28.9**	18.4	0.0
j	1.1	4.6	0.6	1.1	0.6	24.1	0.6	1.1	4.0	**60.3**	1.7
k	11.4	0.0	0.0	2.9	2.9	14.3	0.0	8.6	5.7	45.7	**8.6**

From the results in Table VI, we have the following observations. First, DCNNs, which achieve reasonable results for large-scale image recognition setting, such as the VGG network and AlexNet, are not efficient for facial expression recognition. Second, all the deep features outperform the unlearned features used in the baseline system by a significant margin, which indicates that the deep learning architecture is more robust and applicable for both basic and compound expression classification. Finally, our new LP loss model achieved better performance than the based model and the center loss model. Note that the center loss, which efficiently converges unimodal class, can help to enhance the network performance when recognizing basic emotions, but it failed when applied to compound emotions. These results demonstrate the advantages of the LP loss for multi-modal facial expression recognition, including both basic and compound emotions.

Fig. 9. The distribution of deeply learned features in (a) DCNN without LP loss and (b) DLP-CNN. The locality-preserving loss layer helps the network to learn features with greater discrimination. Moreover, non-neutral expressions that have obvious intensity variations, such as happiness, sadness, fear, surprise and anger, change the intensity continuously and smoothly, from low to high, from center to periphery. Moreover, images with the disgust label, which is the most confused expression, are assembled in the middle. With the neighborhood-preserving character of DLP-CNN, the deep features are able to capture the intrinsic expression manifold structure to a large extent. Best viewed in color.

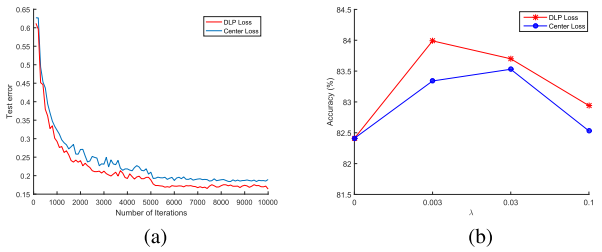


Fig. 10. Comparison of DLP-CNN and center loss in terms of convergence performance and parameter sensitivity. (a) convergence. (b) Accuracy w.r.t λ.

C. Comprehensive Comparisons With Center Loss

As center loss is a special case of DLP-CNN, we further compared DLP-CNN with center loss in terms of time efficiency, convergence performance and parameter sensitivity. For the time efficiency comparison, we evaluated the running time speed using the same settings as those described in Subsection V-B. Specifically, we ran each network for 10 trials over 500 iterations. The running time is 333.18 ± 0.33 ms for center loss and 475.47 ± 2.15 ms for DLP-CNN. According to the setting in [71], the center of each class is directly learned by the network. Considering that the training data size for facial expression recognition is not large, in DLP-CNN, we calculate the k-nearest neighbors for each sample by traversing the whole training set. In this context, DLP-CNN takes more time for each iteration. We have also assessed the convergence performance. Figure 10(a) shows the testing errors of these two methods on basic expression recognition task, which indicates that DLP-CNN has similar convergence speed to that of center loss with improved accuracy during the convergence process. We also

TABLE IX

COMPARISON OF THE RESULTS OF DLP-CNN AND OTHER STATE-OF-THE-ART METHODS ON THE CK+, SFEW 2.0 AND MMI DATABASES. TO VALIDATE THE GENERALIZABILITY OF OUR MODEL, THE WELL-TRAINED DLP-CNN WAS EMPLOYED AS A FEATURE EXTRACTION TOOL WITHOUT FINE-TUNING. (A) CK+. (B) SFEW 2.0. (C) MMI

(a)

Method	Accuracy
CSPL [81]	88.89%
FP+SAE [47]	91.11%
AUDN [41]	92.05%
AURF [41]	92.22%
3DCNN-DAP [42]	92.4%
Inception [52]	93.2%
Dis-ExpLet [44]	95.1%
ESL [65]	95.33%
DLP-CNN (without fine-tuning)	**95.78%**

(b)

Method	Accuracy
DL-GPLVM [19]	24.70%
AUDN [41]	26.14%
STM-ExpLet [43]	31.73%
Inception [52]	47.7%
SFEW third [54]	48.5%
SFEW second [77]	52.29%
SFEW best [33]	52.5%
DLP-CNN (without fine-tuning)	**51.05%**

(c)

Method	Accuracy
3DCNN-DAP [42]	63.4%
DTAGN [30]	70.24%
CSPL [81]	73.53%
AUDN [41]	74.76%
STM-ExpLet [43]	75.12%
F-Bases [63]	75.12%
Inception [52]	77.6%
Dis-ExpLet [44]	77.6%
DLP-CNN (without fine-tuning)	**78.46%**

Fig. 11. Already aligned sample images in CK+, SFEW2.0 and MMI.

checked the sensitivity of parameter λ. Figure 10(b) shows the accuracy of DLP-CNN and center loss, respectively, by varying $\lambda \in \{0, 0.003, 0.03, 0.1\}$ for the basic expression recognition task. The performance of DLP-CNN reaches its maximum when $\lambda = 0.003$ and then decreases as λ continues to increase. This change curve exemplifies the promoting effect of the joint supervision of the softmax loss and the DLP loss when a proper trade-off is chosen. Furthermore, comparison of the performances of DLP loss and center loss shows that DLP-CNN behaves better than center loss as λ varies.

D. Generalizability Tests on DLP-CNN

To assess the generalizability of our well-trained DLP-CNN model to other databases, we employed the model to directly extract fixed-length features of CK+, MMI and SFEW 2.0 without fine-tuning. Already aligned sample images from these three datasets are shown in Figure 11. For the lab-controlled CK+ database, we selected the last frame of each sequence with the peak expression, 309 images in total. During the experiment, we followed the subject-independent experimental principle and performed fivefold cross-validation. For the lab-controlled MMI database, we selected the three peak frames in each sequence for prototypic expression recognition, 528 images in total. Similar to the settings in CK+, we followed the subject-independent experimental principle and performed fivefold cross-validation. For the real-world SFEW 2.0 database, we followed the rule in EmotiW 2015 [12]. "SFEW best [33]", "SFEW second [77]" and "SFEW third [54]" indicate the best single model result of the winner, the runner-up and the second runner-up in EmotiW 2015, respectively. Note that Kim *et al.* [33], Ng *et al.* [54], and Yu and Zhang [77] all trained their models with additional data from SFEW.

From the comparison results in Table IX, we can see that our network can also achieve comparable or even better performance than other state-of-the-art methods, not only for RAF, but also other databases. This indicates that our proposed network can be used as an efficient and effective feature extraction tool for facial expression databases, without a significant amount of time to execute in traditional DCNNs.

VI. CONCLUSIONS AND FUTURE WORK

The main contribution of this paper is the presentation of a new real-world publicly available facial expression database with labeled data from the Internet, based on which we propose a novel optimized algorithm for crowdsourcing and a new locality-preserving loss layer for deep learning. The RAF-DB contains, 1) 29,672 real-world images labeled with different expressions, age range, gender and posture features, 2) a 7-dimensional expression distribution vector for each image, 3) two different subsets: single-label subset, including seven classes of basic emotions; two-tab subset, including twelve classes of compound emotions, 4) the locations of five manually labeled landmark points, and 5) baseline classifier outputs for basic emotions and compound emotions.

For the baseline results, the performances of the frequently used algorithms on RAF-DB, including both shape and appearance features and the SVM classifier, were compared with that of the laboratory-condition databases. The comparison suggests that these methods are unsuitable for expression detection in uncontrolled environments. To solve the problem of real-world expression detection, we tested various deep learning techniques. The proposed method, deep locality-preserving CNN (DLP-CNN), are able to learn more discriminative features for the expression recognition task and help to enhance the classification performance.

We hope that the release of this database will encourage more researches to study the effects of the real-world expression distribution or detection, and we believe that the database will be a useful benchmark resource for researchers to compare the validity of their facial expression analysis algorithms in challenging conditions. In the future, we will attempt to expand the quantity and diversity of our database, especially labels such as fear and disgust, which have relatively few images due to the imbalanced emotion distribution in real-world conditions.

ACKNOWLEDGEMENT

The authors would like to acknowledge Dr. Wenjin Yan and Prof. Xiaotian Fu for employing certified FACS coders and consulting on AU annotation.

REFERENCES

[1] (2010). *Facebook Press Room Statistics*. Accessed: 2010. [Online]. Available: http://www.facebook.com/press/info.php?statistics

[2] T. Bänziger and K. R. Scherer, "Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus," in *Blueprint for Affective Computing: A Sourcebook*. Oxford, U.K.: Oxford Univ. Press, 2010, pp. 271–294.

[3] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[4] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 5562–5570.

[5] V. Bettadapura. (2012). "Face expression recognition and analysis: The state of the art." [Online]. Available: https://arxiv.org/abs/1203.6722

[6] J. Burgess and J. Green, *YouTube: Online Video and Participatory Culture*. Hoboken, NJ, USA: Wiley 2013.

[7] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[8] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image Vis. Comput.*, vol. 24, no. 6, pp. 605–614, 2006.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[10] C. Darwin, P. Ekman, and P. Prodger, *The Expression of the Emotions in Man and Animals*. New York, NY, USA: Oxford Univ. Press, 1998.

[11] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 509–516.

[12] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 423–426.

[13] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014, pp. 647–655.

[14] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[15] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden, "Social media update 2014," Pew Res. Center, Washington, DC, USA, Tech. Rep. 20144, Jan. 2015.

[16] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, pp. 384–392, 1993.

[17] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. Salt Lake City, UT, USA: Consulting Psychologists Press, 2002.

[18] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.

[19] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.

[20] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, pp. 259–275, Jan. 2003.

[21] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, no. 1, pp. 27–38, Jan. 2009.

[22] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*. Berlin, Germany: Springer, 2013, pp. 117–124.

[23] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[24] J. He, J.-F. Hu, X. Lu, and W.-S. Zheng, "Multi-task mid-level feature learning for micro-expression recognition," *Pattern Recognit.*, vol. 66, pp. 44–52, Jun. 2017.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1026–1034.

[26] X. He and P. Niyogi, "Locality preserving projections," in *Proc. NIPS*, vol. 16, 2003, pp. 153–160.

[27] K.-C. Huang, S.-Y. Huang, and Y.-H. Kuo, "Emotion recognition based on a novel triangular facial feature extraction method," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2010, pp. 1–6.

[28] (Dec. 2013). *Face++ Research Toolkit*. [Online]. Available: http://www.faceplusplus.com

[29] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[30] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2983–2991.

[31] S. E. Kahou *et al.*, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[32] S. E. Kahou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.

[33] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jun. 2016.

[34] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, Jan. 2007.

[35] 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.

[36] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 591–600.

[37] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognit. Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[38] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 503–510.

[39] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.

[40] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[41] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.

[42] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proc. Asian Conf. Comput. Vis.* Springer, 2014, pp. 143–157.

[43] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1749–1756.

[44] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets via universal manifold model for dynamic facial expression recognition," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5920–5932, Dec. 2016.

[45] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1805–1812.

[46] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 94–101.

[47] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Proc. Int. Conf. Smart Comput. (SMARTCOMP)*, Nov. 2014, pp. 303–308.

[48] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.

[49] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[50] A. M. Martinez and R. Benavente, "The AR face database," CVC Tech. Rep. 24, Jun. 1998.

[51] D. McDuff, R. El Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected 'in-the-wild,"' in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops (CVPRW)*, Jun. 2013, pp. 881–888.

[52] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[53] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2017.2740923.

[54] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 443–449.

[55] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[56] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[57] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2005, pp. 1–5.

[58] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression recognition using hidden Markov models," *Signal Process., Image Commun.*, vol. 17, no. 9, pp. 675–688, 2002.

[59] X. Peng, Z. Xia, L. Li, and X. Feng, "Towards facial expression recognition in the wild: A new database and deep recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun./Jul. 2016, pp. 93–99.

[60] T. Poell and E. Borra, "Twitter, YouTube, and Flickr as platforms of alternative journalism: The social media account of the 2010 Toronto G20 protests," *Journalism*, vol. 13, no. 6, pp. 695–713, 2012.

[61] J. A. Russell, "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies," *Psychol. Bull.*, vol. 115, no. 1, pp. 102–141, 1994.

[62] A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, "Effective geometric features for human emotion recognition," in *Proc. IEEE 11th Int. Conf. Signal Process.*, Oct. 2012, pp. 623–627.

[63] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Learning bases of activity for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1965–1978, Apr. 2017.

[64] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[65] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh, "Robust representation and recognition of facial emotions using extreme sparse learning," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2140–2152, Jul. 2015.

[66] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[67] Y. Tang. (2013). "Deep learning using linear support vector machines." [Online]. Available: https://arxiv.org/abs/1306.0239

[68] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[69] N. A. Van House, "Flickr and public image-sharing: Distant closeness and photo exhibition," in *Proc. CHI Extended Abstr. Human Factors Comput. Syst.*, 2007, pp. 2717–2722.

[70] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. I-511.

[71] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 499–515.

[72] J. Whitehill and C. W. Omlin, "Haar features for FACS AU recognition," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, Apr. 2006, pp. 96–101.

[73] J. Whitehill, T.-F. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2035–2043.

[74] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[75] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Sep. 2008, pp. 1–6.

[76] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, Apr. 2006, pp. 211–216.

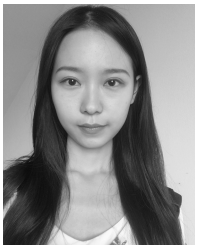
[77] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 435–442.

[78] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[79] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[80] X. Zhang *et al.*, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.

[81] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2562–2569.



Shan Li received the B.E. degree in telecommunication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016. She is currently pursuing the Ph.D. degree in information and telecommunications engineering. Her research interests include facial expression analysis and deep learning.

Weihong Deng received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004 and 2009, respectively. From 2007 to 2008, he was a Post-Graduate Exchange Student with the School of Information Technologies, University of Sydney, Australia, through the support of the China Scholarship Council. He is currently a Professor with the School of Information and Telecommunications Engineering, BUPT. His research interests include statistical pattern recognition and computer vision, with a particular emphasis on face recognition.