Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild

Shan Li, Weihong Deng, and JunPing Du
Beijing University of Posts and Telecommunications
{ls1995, whDeng, junpingd}@bupt.edu.cn

Abstract

Past research on facial expressions have used relatively limited datasets, which makes it unclear whether current methods can be employed in real world. In this paper, we present a novel database, RAF-DB*, which contains about 30000 facial images from thousands of individuals. Each image has been individually labeled about 40 times, then EM algorithm was used to filter out unreliable labels. Crowdsourcing reveals that real-world faces often express compound emotions, or even mixture ones. For all we know, RAF-DB is the first database that contains compound expressions in the wild. Our cross-database study shows that the action units of basic emotions in RAF-DB are much more diverse than, or even deviate from, those of lab-controlled ones. To address this problem, we propose a new DLP-CNN (Deep Locality-Preserving CNN) method, which aims to enhance the discriminative power of deep features by preserving the locality closeness while maximizing the inter-class scatters. The benchmark experiments on the 7-class basic expressions and 11-class compound expressions, as well as the additional experiments on SFEW and CK+ databases, show that the proposed DLP-CNN outperforms the state-of-the-art handcrafted features and deep learning based methods for the expression recognition in the wild.*

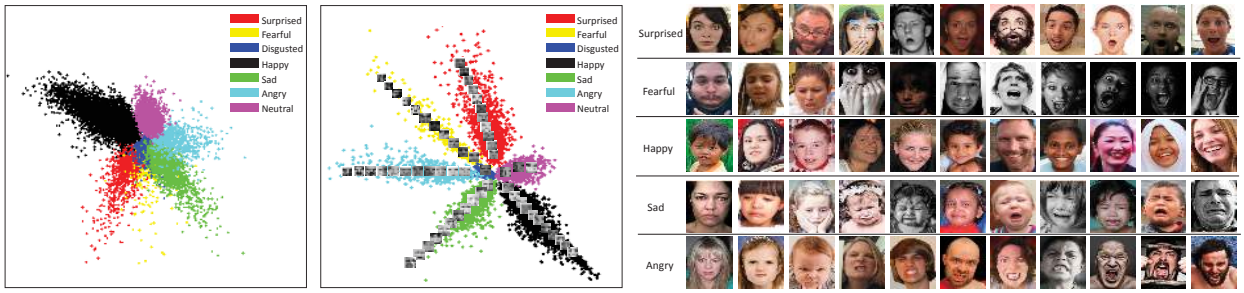
1. Introduction

Millions of images are being uploaded every day by users from different events and social gatherings. There is an increasing interest in designing systems capable of understanding human manifestations of emotional attributes and affective displays. To automatic learn the affective state of face images from the Internet, large annotated databases are required. However, the complexity of annotations of emotion categories has hindered the collection of large annotated databases. On the other side, popular AU coding [12] requires specific expertise to take months to learn and be perfected, hence, alternative solutions are needed. And due to the cultural difference in the way of perceiving facial emotion [13], it is difficult for psychologists to define definite prototypical AUs for each facial expressions. Therefore, it is also worth to study the emotion of social images from the judgments of a large common population, besides from the professional knowledge of a few experts.

In this paper, we propose to study the common expression perception by a reliable crowdsourcing approach. Specifically, our well-trained annotators are asked to label face images with one of the seven basic categories [11], and each face is annotated enough times independently, i.e. about 40 times in our experiment. Then, the noisy labels are filtered by an EM based reliability evaluation algorithm, through which each image can be represented reliably by a 7-dimensional emotion probability vector. By analyzing 1.2 million labels of 29672 great-diverse facial images downloaded from the Internet, these Real-world Affective Faces (RAF)[1] are naturally categorized into two types: basic expression with single-modal distribution and compound emotions with bimodal distribution, an observation supporting a recent ground-breaking finding in the lab-controlled condition [10]. To the best of our knowledge, the real-world expression database RAF-DB is the first large-scale database providing the labels of common expression perception and compound emotions in unconstrained environment.

The cross-database experiment and AU analysis on RAF-DB indicates that AUs of real-world expressions are much more diverse than, or even deviate from, those of lab-controlled ones guided by psychologists. To address this ambiguity of unconstrained emotion, we further propose a novel Deep Locality-preserving CNN (DLP-CNN). Inspired by [17], we develop a practical back-propagation algorithm which creates a locality preserving loss (LP loss) aiming 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

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



(a) DCNN without LP loss (b) DLP-CNN (c) Amplified face images in color from Subfigure (b)

Figure 1. The distribution of deeply learned features in (a) "DCNN without LP loss" and (b) "DLP-CNN". As can be seen, locality preserving loss layer helps the network to learn features with more discrimination. Moreover, it can be clearly seen that non-neutral expressions which 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. And images with Disgust label, which is the most confused expression, are assembled in the middle. With the neighborhood preserving character of DLP-CNN, the deep feature seems to be able to capture the intrinsic expression manifold structure to a large extent. Best viewed in color.

class to become compact, and thus the discriminative power of the deeply learned features can be highly enhanced. Moreover, locally neighboring faces tend to share similar emotion intensity by using DLP-CNN, which can derive the discriminative deep feature with smooth emotion intensity transition. Figure 1 (b) shows the resulting 2-dimensional deep features learnt from our DLP-CNN model, where we attach example face images with various intensity in different expression classes.

Extensive experiments on RAF-DB and other related databases show that the proposed DLP-CNN outperforms other state-of-the-art methods. Moreover, the activation features trained on RAF-DB can be re-purposed 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).

2. Related Work

2.1. Expression image datasets

Facial expression recognition largely relies on well-defined databases, however, several limitations exist.

Many available databases were produced in tightly controlled environments without diversity on subjects and conditions. Subjects in them were taught to act expressions in a uniform way. Besides, the majority of current databases only include six basic categories or less. However, images captured in real-life scenarios often present complex, compound or even ambiguous emotions rather than simple and prototypical ones [3]. What's more, labelers in these databases are too few, which would reduce the reliability and validity of the emotion labels.

We then focus on discussing image databases with spontaneous expressions. SFEW 2.0 [7] contains 700 images extracted from movies, and images were labelled by two independent labelers. The database covers unconstrained facial expressions, varied head poses, large age range, occlusions, varied focus, different resolution of face. FER-2013 [16] contains 35887 images collected and labelled using the Google image search API. Cropped images are provided in 48×48 pixels and converted to grayscale. BP4D-Spontaneous [47] contains plenty of images from 41 subjects revealing a range of spontaneous expressions elicited through eight tasks. However, the database organization were lab-controlled. AM-FED [30] is collected in real world with sufficient samples, however, without specifical emotion labels, it's more suited for researches on AUs. EmotioNet [1] is a large database of one million facial expression images in the wild created by an automatic AU detection algorithm. Unlike these databases, RAF-DB simultaneously satisfies multiple requirements: sufficient data, various environments, group perceiving on facial expressions and data labels with the least noise.

2.2. The framework for expression recognition

Facial expression analysis can be generally divided into three main parts [14]: face aquisition, facial feature extraction and facial expression classification.

In face aquisition 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 generally categorized into two groups: Appearance-based methods [29] and AU-based methods [42]. The former uses common feature extraction methods such as LBP [38], Haar [44]. The latter recognizes expression by detecting AUs. Feature classification is performed in the last stage. The commonly used methods include SVM, nearest neighbor, LDA, DBN and decision-level fusion on these classifiers [46]. The extracted facial

expression information is either classified as a set of facial actions or a particular basic emotion [34]. Most focus on the latter and is based on Ekman's theory of six basic emotions [12]. 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 is therefore not universal enough to generalize expressions displayed on human face [37].

2.3. 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, which achieve state-of-the-art results. So far, there have been a few deep neural networks used in facial expression recognition due to the lack of sufficient training samples. In ICML 2013 competition [16], the winner [41] was based on Deep Convolutional Neural Network (DCNN) plus SVM. In EmotiW 2013 competition [6], the winner [19] combined modality specific deep neural network models. In EmotiW 2015 [8], more competitors have tried deep learning methods: transfer learning was used to solve the problem of small database in [32], hierarchical committee of multi-column DCNNs in [20] gained the best result on SFEW database, LBP features combined with DCNNs structure were proposed in [22]. In [24], AU-aware Deep Networks (AUDN) was proposed to learn features with the interpretation of facial AUs. In [31], a DCNN with inception layers was proposed to gain comparable results.

3. Real-world Expression Database: RAF-DB

3.1. Creating RAF-DB

Data collection. At the very beginning, the 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 (for example: smile, giggle, cry, rage, scared, frightened, terrified, shocked, astonished, disgust, expressionless) to pick out images that were related with the six basic emotions plus the neutral emotion. At last, a total of 29672 real-world facial images are presented in our database. Figure 2 shows the pipeline of data collection.

Database annotation. Annotating nearly 30000 images of expression is an extremely difficult and time-consuming task. Considering the compounded property of real-world expressions, multiple views of images' expression state should be collected from different labelers. We therefore employed 315 annotators (students and staffs from universities) who have been instructed with one-hour tutorial of

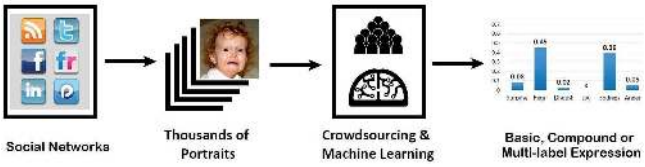


Figure 2. Overview of construction and annotation of RAF-DB.

psychological knowledge on emotion for an online facial expression annotation assignment, where they were asked to classify the image into the most apparent one from seven classes. We developed a website for RAF-DB annotation, which shows each image with exclusive attribute options. Images were randomly and equally assigned to each labeler, ensuring that there were no direct correlation among the images labeled by one person. And each image was assured to be labeled by about 40 independent labelers. After that, a multi-label annotation result is obtained for each image, i.e., a seven dimensional vector that each dimension indicates the votes of relevant emotion.

Metadata. The data is provided with precise locations and size of the face region, as well as the manually located five landmark points (the central of two eyes, the tips of the nose and two corners of the mouth) on the face. Besides, an automatic landmark annotation mode without manual label is included: 37 landmarks were picked out from the annotation results provided by Face++ API [18]. 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% remains unsure. 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.

Reliability estimation. Due to subjectivity and varied expertise of labelers and wide ranging levels of images' difficulty, there were some disagreements among annotators. To get rid of noisy labels, motivated by [45], a Expectation Maximization (EM) framework was used to assess each labeler's reliability.

Let $D = \{(x_j, y_j, t_j^1, t_j^2, ..., t_j^R)\}_{j=1}^n$denote a set of n labeled inputs, where y_j is the gold standard label (hidden variable) for the jth samples x_j , $\mathrm{t}_j^i \in \{1, 2, 3, 4, 5, 6, 7\}$ is the corresponding label given by the ith annotator. The correct probability of t_j^i are 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 images, α_i is the reliability of ith annotators.

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

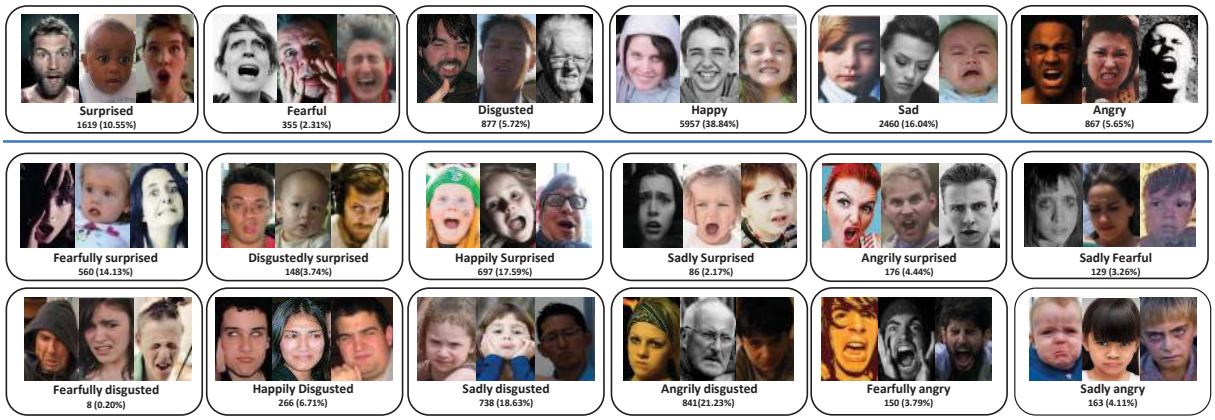


Figure 3. Examples of six-class basic emotions and twelve-class compound emotions from RAF-DB. Detailed data distribution of RAF-DB has been attached to each expression classes.

labels:

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

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.

Subset Partitions. Let $G_j = \{g_1, g_2, ..., g_7\}$ denotes 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 $k \in \{1,2,3,4,5,6,7\}$ refer to surprise, fear, disgust, happiness, sadness, anger and neutral, respectively. We then divided RAF-DB into different subsets according to the 7-dimensional ground truth. For Single-label Subset, we first calculated the mean distribution value $g_{mean} = \sum_{k=1}^{7} g_k/7$ for each image, then picked out label k w.r.t. $g_k > g_{mean}$ as the valid label. Images who have single valid label are classified into Single-label Subset. For Two-tab Subset, the partition rule is similar. The only difference is that we took out images with neutral label before partition. Figure 3 exhibits specific samples of 6-class basic emotions and 12-class compound emotions.

3.2. CK+ and RAF Cross-Database Study

We then conducted a CK+ [26] and RAF cross-database study to explore the specific difference between expression-

Algorithm 1 Label reliability estimation algorithm.

Input: Training set $D = \{(x_j, t_j^1, t_j^2, ..., t_j^R)\}_{j=1}^n$
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:

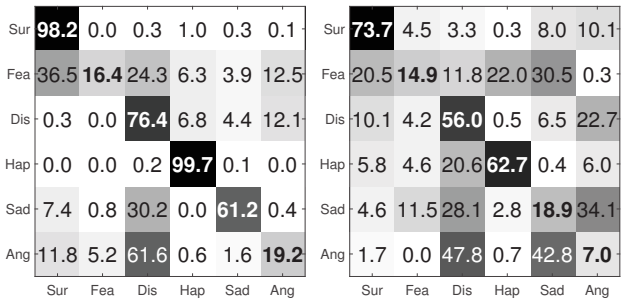
$$\alpha_i := \arg\max_{\alpha_i} \sum_j \sum_{y_j} Q_j(y_j) \ln \frac{p(t_j,y_j|\alpha_i,\beta_j)}{Q_j(y_j)}$$

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

s of real-world affective face and the lab-controlled posed face guided by psychologist. Here, "cross-database" means we use all of the images from one database for training and the images from the other for testing. In order to eliminate the bias caused by different training size, the single-tab subset of RAF-DB has been sub-sampled for experiment to balance the size of two databases.

To ensure the generalization capabilities of the classifiers, we applied support vector machine for classification and tried HOG descriptor [5] for representation. Specifically, original images were first aligned to the size of 100×100. Then, we got a 4000-dimensional HOG feature vector per aligned image. Finally, SVM with RBF kernel implemented



(a) RAF → CK+ (b) CK+ → RAF

Figure 4. Confusion matrixes for cross-database experiments using HOG features. The true labels (training data) are on the vertical axis, the predicted labels (test data) are on the horizontal axis.

by LibSVM [4] was applied for classification. Parameters were optimized using grid search.

We then performed a cross-database experiment based on six-class expression. Multiclass support vector machine (mSVM) and confusion matrix were used as the classification method and the assessment criteria respectively. Figure 4 shows the results of this experiment.

Analyzing the diagonal of these two matrixes, we can see that surprise, happiness and disgust are the top three that have the highest recognition rates in both cases. This result is in line with many single database tests based on CK+, such as [26], [35] and [38]. After calculating the average of the diagonals, Matrix I was detected with 62% accuracy while Matrix II with only 39%, which indicates that data collected from real world is more multiple and effective than lab-controlled one. This is particularly evident in the expression of sadness, then happiness and surprise. Besides, anger and disgust are usually confused with each other in both cases, which conforms to the survey in [2].

In order to explain the phenomena above, a more detailed research must be conducted to find out the specifical differences of each expression between these two databases. Therefore, a facial action coding system (FACS) analysis has been employed. FACS was first presented in [12], where the changes on facial behaviors are described by a set of action units (AUs). AUs of sub-sampled images in RAF-DB were first labeled by our FACS coders. We then quantitatively analyzed the AU presence for different emotions in CK+ and RAF. Some examples from CK+ and RAF are shown in Figure 5. Besides, probabilities of AUs' occurrence for each expression from sub-sampled images in RAF-DB have been shown in Table 1.

4. Deep Locality-Preserving Feature Learning

Besides the "in-the-wild" difficulties such as variable lighting, poses and occlusions, real-world affective faces at least pose two challenges that demand new algorithms to address. First, as indicated by our cross-database study, real world expression 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 amount of real-world affective faces express compound, or even multiple emotions. So traditional hand-engineered representations which perform well on the lab-controlled databases are no longer suitable for expression recognition tasks in the wild.

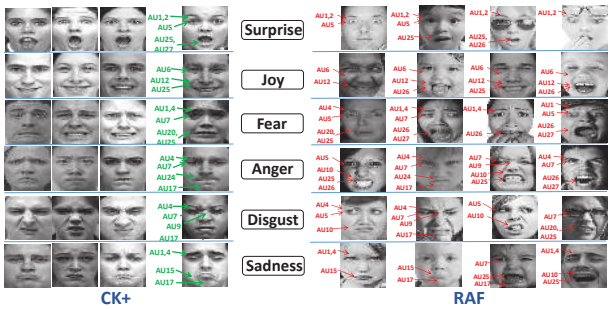


Figure 5. Comparison of six basic emotions from CK+ and RAF. It's evident that expression AUs in RAF are more diverse than those in CK+.

Table 1. Probabilities of AUs' occurrence for each expression in RAF-DB

(%)	AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU15	AU 17	AU20	AU25	AU 26	AU27
Sur	97	97		84									98	53*	
Fea	78	42	74	79		50						30*		61*	43*
Dis			51				34*	89*			82	26	55*		
Hap					98				85				97	23	
Sad	88		84							21*	54		49*		
Ang			96	72*		94		36			87		79*	72*	

The empty 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).

Nowadays, DCNN has been proved to outperform hand-crafted features on lager-scale visual recognition tasks. Nevertheless, conventional DCNN uses only the softmax loss layer to supervise the training process. The softmax layer helps keeping the deeply learned features of different classes separable, however, still remains serious intra-class variation. On the contrary, facial expressions in real world show significant intra-class difference on account of varied occlusions, illuminations, resolutions and head positions. What's more, individual variation can also lead to big difference for the same category expression, for example, laugh v.s. smile. Hence, we proposed a novel DLP-CNN to address the ambiguity and multi-modality of real-world facial expressions. In DLP-CNN, we added a new supervised layer on the fundamental architecture shown in Table 2, 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 make the local neighborhoods within each class as

Table 2. 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	-	2000	-	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	

compact as possible. To formulate our goal:

$$\min_{W} \sum_{i,j} S_{ij} ||x_i - x_j||_2^2 \tag{1}$$

where W is the network parameters, and the matrix S is a similarity matrix. The deep feature $x \in \mathbb{R}^d$ denotes Deep Convolutional activation features (DeCaf) [9] taken from the final hidden layer, i.e., just before the softmax layer that produces the class prediction. A possible way of defining S is as follows.

$$S_{ij} = \begin{cases} 1, \ x_j \text{ is among } k \text{ nearest neighbors of } x_i \\ \quad \text{or } x_i \text{ is among } k \text{ nearest neighbors of } x_j \\ 0, \ \text{otherwise} \end{cases} \tag{2}$$

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

This formulation effectively characterizes the intra-class local scatters. Note that x_i should be updated as the iterative optimization of the CNN. To compute the summation of the pairwise distance, we need to take 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 locality preserving loss function of x_i is defined as follow:

$$L_{lp} = \frac{1}{2} \sum_{i=1}^{n} ||x_i - \frac{1}{k} \sum_{x \in N_k\{x_i\}} x||_2^2 \tag{3}$$

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

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

$$\frac{\partial L_{lp}}{\partial x_i} = x_i - \frac{1}{k} \sum_{x \in N_k\{x_i\}} x \tag{4}$$

In this manner, we can perform the update based on mini-batch. Note that, the recently proposed center loss [43] can be considered as a special case of the locality preserving loss, if $k = n_c - 1$ (n_c is the number of the training samples in class c to which x_i belong). While center loss simply pulls the samples to a single centroid, the proposed locality preserving 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 locality preserving loss which characterizes the local scatters within class, to train the CNNs for discriminative feature learning.

The objective function is formulated as follow: $L = L_s + \lambda L_{lp}$, where L_s denotes the softmax loss and L_{lp} denotes the locality preserving loss. The hyper parameter λ is used to balance the two loss functions. Algorithm 2 summarizes the learning process in the deep locality preserving CNN.

Algorithm 2 Optimization algorithm of DLP-CNN.

Input: Training data $\{x_i\}_{i=1}^n$,
n is the size of mini-batch
Output: Network layer parameters W
Initialize: $t = 0$
Network learning rate μ, hyper parameter λ, Network layer parameters W, softmax loss parameters θ, neighboring nodes k.
Repeat:
1: $t = t + 1$
2: Computer the center of 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: Computer 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

5. Baseline System

To facilitate translating the research from laboratory environments to the real world, we performed two challenging benchmark experiments on RAF-DB: 7-class basic expression classification and 11-class compound expression classification, and presented affiliated baseline algorithms and performances. We also conducted comparative experiments on two small and popular datasets, CK+ and JAFFE [28].

We followed up the experimental setup in cross-database experiments, and tried LBP [33], HOG [5] and Gabor [23] features. The LBP descriptor applied the 59-bin $LBP_{8,2}^{u2}$ operator, and then concatenated the histograms from 10×10 pixel cells, generating a 5,900 dimensional feature vector. The HOG feature used this shape-based segmentation dividing the image into 10×10 pixel blocks of four 5×5 pixel cells with no overlapping. By setting 10 bins for each histograms, we extract a 4000-dimensional HOG feature vector for each image. For Gabor wavelet, we used a bank of

Table 3. Basic expression class performance comparison of CK+, JAFFE and RAF along with Compound expression performance of RAF, based on LBP, HOG and Gabor descriptors, and SVM, LDA+kNN classification. The metric is the mean diagonal value of the confusion matrix.

		basic			compound
		CK+	JAFFE	**RAF**	**RAF**
mSVM	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	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**

40 Gabor filters at five spatial scales and eight orientations. The downsample image's size was set to 10*10, yielding 4000-dimensional features.

In order to objectively measure the performance for the followers entries, we split the dataset into a training set and a test set with the idea of five-fold cross-validation, which means the size of training set is five times larger than test set, and expressions in both sets have a near-identical distribution. Considering expressions in the wild have imbalanced distribution, the accuracy metric which is especially sensitive to bias and no longer effective for imbalanced data [15], is no longer used in RAF. Instead, we use the mean diagonal value of the confusion matrix as the ultima metric.

Basic emotions. In this experiment, seven basic emotion classes were detected using the whole 15339 images from the single-label subset. The best classification accuracy (output by SVM) was 72.71% for LBP, 74.35% for HOG, and 77.28% for Gabor. Results declined to 55.98%, 58.45% and 65.12% respectively when using the mean diagonal value of the confusion matrix as metric. To assess the reliability of the basic emotion labels, we also assigned a uniform random label to each sample, which we call a naive emotion detector. And the best result for the naive classifier was 16.07% when using Gabor feature, which is much lower than the former value.

For comparison, we employed the same methods on CK+ with person-independent 5-fold cross-validation and JAFFE with leave-one-subject-out strategy. The results shown in Table 3 certify that expressions in real world are more difficult for recognition and the current common methods which perform well on the existing databases cannot solve the expression recognition problem in the challenging real-world condition.

To evaluate effectiveness of different classifiers, we have also trained LDA with nearest neighbor (NN) classification. We found that LDA+NN were inferior to mSVM obviously when training on RAF, a extremely large database. Nevertheless, it performed better when training on small-scale datasets (CK+ and JAFFE), even outperformed mSVM in some cases. Concrete results can be viewed in Table 3.

Compound emotions. For compound emotions classification, we got rid of fearfully disgusted emotion as it's too few, leaving 11 classes of compound emotion, 3954 in total. The best classification accuracy (output by SVM) was 45.51% for LBP, 51.89% for HOG, and 53.54% for Gabor. Results declined to 28.84%, 33.65% and 35.76% respectively when using the mean diagonal value of the confusion matrix as metric. Again, to demonstrate the reliability of the compound emotion labels, we computed the baseline for the naive emotion detector, which declined to 5.79% when using Gabor feature.

As expected, the overall performance dropped significantly when more expressions are involved for classification. The significantly lower results compared to that of basic emotions indicate that compound emotions are more difficult to detect and new methods should be invented to solve this problem. Besides the multi-modality, lack of training samples of compound expressions from real world is another great technical challenge.

6. Deep Learning System

Nowadays, deep learning has been applied to lager-scale visual recognition tasks and perform exceedingly well with lager amounts of training data. However, fully-supervised deep models are easy to be overfitting on facial expression recognition task due to the insufficient training samples for the model learning. Therefore, most deep learning frameworks employed on facial expression recognition [22, 32, 36] are base on pre-trained models. These pre-trained models, such as VGG network [40] and AlexNet [21], are initially designed for face recognition, which are short of discrimination ability of expression characteristic. So in this paper, we directly trained our deep learning system on the big enough self-collected database RAF from scratch, without using other databases.

When conducting experiments, we followed the same dataset partition standards, image processing methods and classification methods as in the baseline system. Related researches [9, 39] have proved that well-trained deep convolutional network can work as a feature extraction tool with generalization ability for the classification task. Following up this idea, we first trained each DCNNs for basic emotion recognition task, and then directly used the already trained DCNN models to extract deep features for both basic and compound expressions. 2000-dimensional deep features learnt from raw data were extracted from the penultimate fully connected layer of the DCNNs and then classified by SVM.

Table 4. 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**	**Average**
mSVM	VGG	68.52	27.50	35.13	85.32	64.85	66.32	59.88	**58.22**	**31.63**
	AlexNet	58.64	21.87	39.19	86.16	60.88	62.31	60.15	**55.60**	**28.22**
	baseDCNN	70.99	52.50	50.00	92.91	77.82	79.64	83.09	**72.42**	**40.17**
	center loss	68.52	53.13	54.05	93.08	78.45	79.63	83.24	**72.87**	**39.97**
	DLP-CNN	71.60	52.15	62.16	92.83	80.13	81.16	80.29	**74.20**	**44.55**
LDA	VGG	66.05	25.00	37.84	73.08	51.46	53.49	47.21	**50.59**	**16.27**
	AlexNet	43.83	27.50	37.84	75.78	39.33	61.70	48.53	**47.79**	**15.56**
	baseDCNN	66.05	47.50	51.35	89.45	74.27	76.90	77.50	**69.00**	**28.23**
	center loss	64.81	49.38	54.05	92.41	74.90	76.29	77.21	**69.86**	**27.33**
	DLP-CNN	77.51	55.41	52.50	90.21	73.64	74.07	73.53	**70.98**	**32.29**

From the results in Table 4, we have the following observations. First, DCNNs which achieve quite reasonable results for large-scale image recognition setting, such as VGG network and AlexNet, are not efficient for facial expression recognition. Second, all of the deep features learnt on RAF-DB outperform the unlearned features used in the baseline system by a significant margin, which indicates that deep learning architecture is more robust and applicable for both basic and compound expression. At last, our new locality preserving loss model achieves better performance than the based one and the center loss one. Note that, the center loss, which efficiently converges unimodal class, can help enhance the network performance on basic emotion, but it fails on compound emotion. This shows the advantage of the locality preserving loss on multi-modal facial expression recognition, including both basic and compound one.

To see the generalization ability of our well-trained DLP-CNN model on other databases, we then employed it to directly extract fixed-length feature of CK+ and SFEW 2.0 without finetune. For the lab-controlled databases CK+, we followed the experimental principle in the baseline system. For the real-world database SFEW 2.0, we followed the rule in EmotiW 2015 [8], and the "SFEW best" is the result of the single best model used in the winner [20] of EmotiW 2015. Note that, in [20], the Authors trained their model with extra data from SFEW. From the comparison results in Table 5, 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.

Table 5. Comparison results of DLP-CNN and other state-of-the-art deep learning methods on CK+ and SFEW 2.0.

	AUDN [25]	FP+SAE [27]	[31]	SFEW best [20]	**DLP-CNN** (without finetune)
CK+	93.70	91.11	93.2	–	**95.78**
SFEW 2.0	30.14	–	47.7	52.5	**51.05**

7. Conclusions and Future Work

The main contribution of this paper is presenting a novel optimized algorithm for crowdsourcing and a new locality preserving loss layer for deep learning, based on a real-world publicly available facial expression database RAF-DB. The optimized algorithm helps to keep the best annotated results from labelers. The new DCNN can learn more discriminative feature for expression recognition task. The RAF-DB contains, 1) 29672 real-world images labeled for different expressions, age range, gender and posture feature, 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) locations of five manually accurate detect landmark points, 5) baseline classifier outputs for basic emotions and compound emotions. We hope that the release of this database will encourage more researches on the effect of real-world expression distribution or detection and be a useful benchmark resource for researchers to compare the validity of their facial expression analysis algorithms in challenge conditions.

8. Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Project 6157306861471048, 61375031 and 61532006), Beijing Nova Program under Grant No. Z161100004916088, the Fundamental Research Funds for the Central Universities under Grant No. 2014ZD03-01, and the Program for New Century Excellent Talents in University(NCET-13-0683).

References

[1] 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 *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR16), Las Vegas, NV, USA*, 2016.

[2] V. Bettadapura. Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*, 2012.

[3] J. C. Borod. *The neuropsychology of emotion*. Oxford University Press New York, 2000.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE, 2011.

[8] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014.

[10] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[11] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[12] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[13] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[14] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[15] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009.

[16] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural information processing*, pages 117–124. Springer, 2013.

[17] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, volume 16, 2003.

[18] M. Inc. Face++ research toolkit. www.faceplusplus.com, Dec. 2013.

[19] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013.

[20] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, pages 1–17, 2016.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 503–510. ACM, 2015.

[23] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE Transactions on*, 11(4):467–476, 2002.

[24] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[25] M. Liu, S. Li, S. Shan, and X. Chen. Au-inspired deep networks for facial expression feature learning. *Neurocomputing*, 159:126–136, 2015.

[26] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[27] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In *Smart Computing (SMARTCOMP), 2014 International Conference on*, pages 303–308. IEEE, 2014.

[28] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (jaffe) database. 1998.

[29] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1357–1362, 1999.

[30] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 881–888. IEEE, 2013.

[31] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[32] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 443–449. ACM, 2015.

[33] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[34] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, Dec 2000.

[35] M. Pardàs and A. Bonafonte. Facial animation parameters extraction and expression recognition using hidden markov models. *Signal Processing: Image Communication*, 17(9):675–688, 2002.

[36] X. Peng, Z. Xia, L. Li, and X. Feng. Towards facial expression recognition in the wild: A new database and deep recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 93–99, 2016.

[37] J. A. Russell. Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102, 1994.

[38] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[39] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[42] Y. I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, Feb 2001.

[43] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[44] J. Whitehill and C. W. Omlin. Haar features for facs au recognition. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, page 5–pp. IEEE, IEEE, 2006.

[45] 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 *Advances in neural information processing systems*, pages 2035–2043, 2009.

[46] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[47] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.