

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/136936>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Human Emotion Distribution Learning from Face Images using CNN and LBC Features

Abeer Almowallad
Dept. of Computer Science
University of Warwick
Coventry, UK
Abeer.Almowallad@warwick.ac.uk

Victor Sanchez
Dept. of Computer Science
University of Warwick
Coventry, UK
V.F.Sanchez-Silva@warwick.ac.uk

Abstract—Human emotion recognition from facial expressions depicted in images is an active area of research particularly for medical, security and human-computer interaction applications. Since there is no pure emotion, measuring the intensity of several possible emotions depicted in a facial expression image is a challenging task. Previous studies have dealt with this challenge by using label-distribution learning (LDL) and focusing on optimizing a conditional probability function that attempts to reduce the relative entropy of the predicted distribution with respect to the target distribution, which leads to a lack of generality of the model. In this work, we propose a deep learning framework for LDL that uses convolutional neural network (CNN) features to increase the generalization of the trained model. Our framework, which we call EDL-LBCNN, enhances the features extracted by CNNs by incorporating a local binary convolutional (LBC) layer to acquire texture information from the face images. We evaluate our EDL-LBCNN framework on the s-JAFFE dataset. Our experimental results show that the EDL-LBCNN framework can effectively deal with LDL for human emotion recognition and attain a stronger performance than that of state-of-the-art methods.

Index Terms—Human emotion recognition, label distribution learning, local binary convolutional, deep learning, CNN.

I. INTRODUCTION

Facial emotions are essential elements in human communication as they can help to understand the intentions of others. In general, people convey emotional states, e.g., happiness, sadness, or anger, by using facial expressions. Indeed, facial expressions are one of the main elements, among several nonverbal components such as hands gestures and body movements, that carry emotional meaning in human communication. Therefore, emotion classification from face images has been an interesting research topic over the past decades in computer vision, with applications in human-computer interaction, medicine [1] and security [2].

There are many state-of-the-art works [3]–[8] on classifying the six basic human emotions (happiness, sadness, fear, anger, surprise, and disgust) using face images. These works use either single-label learning (SLL) [5]–[8] to predict a single class, or emotion, or multi-label learning (MLL) [3], [4] to classify an image into multiple classes. According to Zhou et al. [9], however, there is no pure human emotion but rather a combination of all emotions that may manifest with different degrees, or intensities. For example, a person may be happy

and surprised at the same time but with different degrees. This calls for a different set of solutions that differ from those based on SLL or MLL [3]–[8].

To accurately detect the degree of all emotions depicted in a single facial expression image, we adopt label-distribution learning (LDL). This type of learning aims to detect the degree of all classes present in each sample. Previous works based on LDL [9], [10] use the maximum entropy (ME) model, which limits the generality of the learning algorithms, since ME models explicitly impose a hypothesis depending on the data, and generalization requires no such hypothesis. Moreover, ME models require an appropriate function to measure the similarity between predicted and target distributions. Such a function may be highly complex, which also affects the model’s generalization capabilities [11].

In this work, we propose a deep-learning framework to transfer the problem of classifying individual and multiple human emotions from facial expression images (SLL or MLL) to measuring the distribution of mixed emotions (LDL). Our framework uses a convolutional neural network (CNN) to account for the spatial information of images in the form of features. To enhance these features, our framework uses another set of features computed by a local binary convolutional (LBC) layer [12]. This LBC layer is based on the concept of local binary patterns (LBP) [13], [14], which have been shown to be effective in extracting texture information from face images. Our framework then exploits this texture information to correctly predict the distribution of emotions depicted in facial expression images. We call our framework Emotion Distribution Learning by LBC and CNN layers (EDL-LBCNN).

Performance evaluations on the s-JAFFE dataset show that the EDL-LBCNN framework can accurately learn the distribution of six basic emotions (happiness, sadness, fear, anger, surprise, and disgust) from facial expression images. Compared to baseline deep-learning methods based only on CNN features and those based on a combination of CNN features and face landmarks, the EDL-LBCNN framework achieves the best performance.

The rest of the paper is organized as follows. Section II reviews related work. In Section III, we explain our EDL-LBCNN framework in detail. Section IV presents our exper-

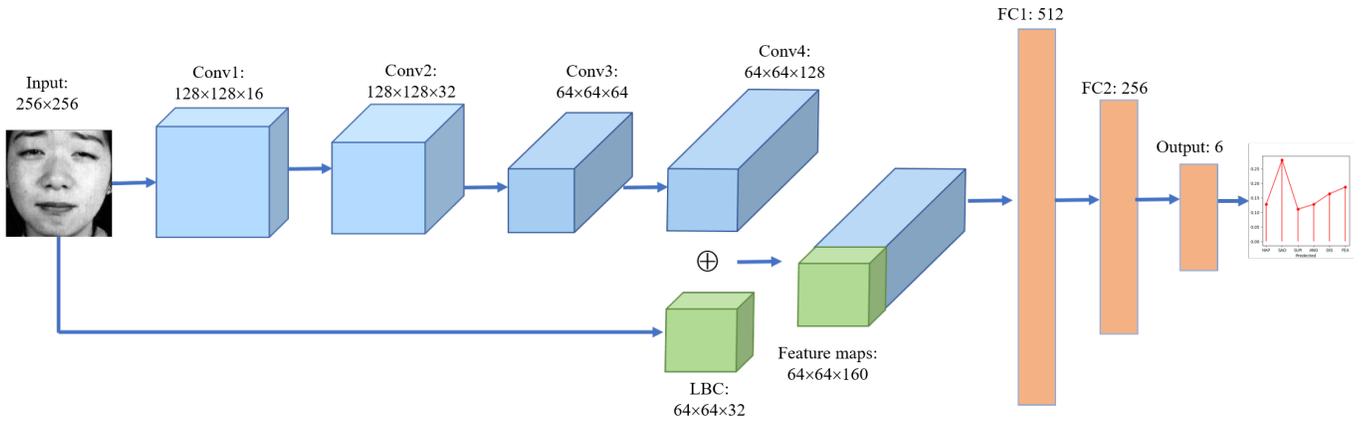


Fig. 1. The architecture of the proposed EDL-LBCNN framework. It has two streams, the first stream is a 4-layer CNN (top part in blue, while the second stream is a single LBC layer (lower part in green). The feature maps extracted by the two streams are concatenated and used as input to two FC layers. The Softmax function is applied to the output layer, while KL Loss is used for training.

iments and analyzes the results. Finally, our conclusions are drawn in Section V.

II. RELATED WORK

Human emotion recognition algorithms from face images may use SLL, MLL, or LDL. In this paper, we focus on reviewing related works on predicting the degree of mixed human emotions from face images, specifically those methods that use deep learning and LDL.

LDL is a learning method proposed by Geng et al. [15] that maps an instance to a distribution, which is a description of the degree or intensity of all the classes present in the dataset. LDL relies on a number of algorithms categorized into three strategies, as explained in [15]: 1) Problem Transformation (PT), 2) Algorithm Adaptation (AA), and 3) Specialized Algorithm (SA). PT transforms an LDL problem into an SLL problem by generating a single-label dataset. For each image in the dataset, c examples are generated for a total of c classes. Each example is assigned a target value equal to the class degree. Consequently, any SLL algorithm can be used on the resulting dataset, e.g., PT-SVM [16] or PT-Bayes [17]. AA takes an existing learning algorithm that can be naturally extended to deal with label distributions, such as AA- K -NN [18] and backpropagation neural network (AA-BP) [19]. The SA strategy considers algorithms that solve the LDL problem by optimising an energy function based on the maximum entropy to learn a label distribution, such as SA-IIS and SA-BFGS in [15].

Zhou et al. propose Emotion Distribution Learning (EDL) [9], which is based on the SA strategy and measures the degree of each emotion from a single facial expression image. The authors base their method on Jeffreys divergence to measure the distance between two distributions and add a correlation coefficient between two emotions to consider the relationship among relevant emotions. Li et al. exploit label correlations in a global manner under the hypothesis that correlations are shared by all instances [10]. EDL-LRL (Emotion Distribution

Learning by exploiting Low-Rank label correlations Locally) [10] is later proposed by Ren et al. to improve the performance of EDL by considering local-label correlations rather than global-label correlations. However, EDL-LRL uses the same SA strategy, which may affect the generalization of the model. Moreover, these methods only use the correlations of relevant labels and not visual representations of images. Different from these algorithms, our proposed EDL-LBCNN framework uses features that account for visual representations of emotions.

To measure the distribution of mixed emotions using visual information, Peng et al. [20] develop a method called CNNR to predict human reactions to a photo by using LDL with CNN features. Their method, which trains a regressor for each emotion to predict a real value, outperforms SVR (Support Vector Regression), proposed in the same study. Another method that uses LDL with CNN features to analyze emotions from facial expression images is proposed by Yang et al. in [21]. Specifically, the authors propose a multi-task framework by jointly optimizing classification and distribution prediction to understand the ambiguity and relationship among emotions.

More recently, several methods based on LDL and CNN features have been proposed to solve different problems using face images, such as age estimation and facial attractiveness estimation. In [22], Yang et al., propose SEU-NJU to estimate the age distribution of an individual based on their face images. Their method uses two streams to extract features, one is pre-trained on the VGG-16 network while the other stream is a CNN architecture with four convolutional layers, four max pooling layers, and two fully connected (FC) layers. This second stream, which is called DCNN, is trained on different types of input images, i.e., RGB, HSV, single channel, gray-level, and Sobel-level images, to increase the capacity to correctly estimate the age. The distribution-based loss function, Kullback-Leibler (KL) divergence, and the Softmax loss function, help the network to exploit the uncertainty information from the distribution values associated with the face images. In [23], Liu et al. propose a method to estimate

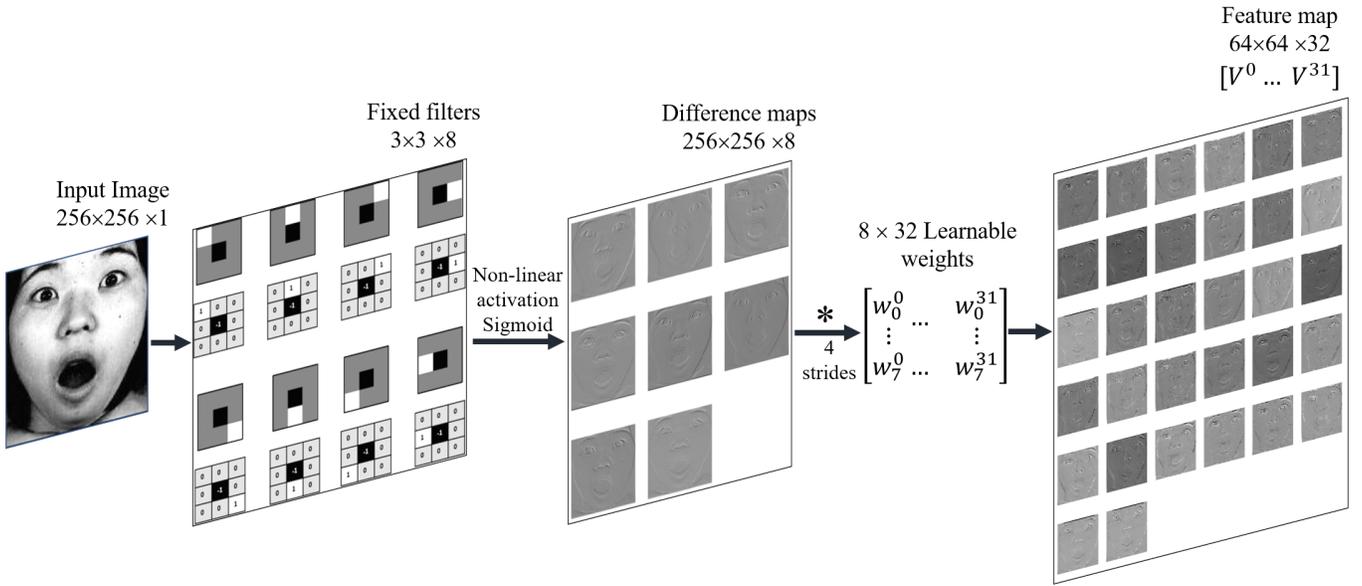


Fig. 2. The LBC layer architecture comprises two main parts. The first part convolves the input image with eight fixed 3×3 filters to produce eight difference maps in a distinct manner. The second part uses 8×32 learnable parameters to output 32 feature maps. Each feature map merges the eight difference maps in a distinct manner.

the degree of attractiveness of an individual based on their face image. They use LDL with CNN features to significantly enhance the generalization ability of their model, which is trained on a small dataset. Their method consists of a fusion of high-level features extracted from a pre-trained ResNet50 network and low-level features, which are extracted in the form of facial landmarks. Euclidean distance and KL divergence are used to define the loss function of their method. Our proposed EDL-LBCNN framework is similar to theirs, as we also use CNN features. However, we incorporate texture information as extracted by an LBC layer.

III. THE EDL-LBCNN FRAMEWORK

Let X be the data space and $Y = \{y_1, y_2, \dots, y_c\}$ the complete set of c labels. The label distribution associated with an sample x_i , where $x_i \in X$, is $D_i = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \dots, d_{x_i}^{y_c}\}$, which results in the dataset $G = \{(x_1, D_1), (x_2, D_2), \dots, (x_n, D_n)\}$, where n is the number of images, $d_{x_i}^{y_j}$ denotes the degree (intensity) of label $y_j \in Y$ in the instance x_i , where $i = 1, 2, \dots, n$, $j = 1, 2, \dots, c$, and $\sum_j d_{x_i}^{y_j} = 1$. In other words, each image in the dataset G is associated with a vector, D_i , of values $\in [0, 1]$. The images in G are used as input to our proposed framework to predict a distribution \hat{D}_i for each x_i , where \hat{D}_i should be very similar to the ground truth distribution, D_i .

Our EDL-LBCNN framework comprises two streams: the first stream has four convolutional layers, while the second has a single LBC layer, as depicted in Fig. 1. The two streams use grayscale images of size 256×256 to extract features. These features are concatenated and used as the input to two FC layers. The framework has six neurons in the output layer,

one for each of the six basic emotions: happiness, sadness, fear, anger, surprise, and disgust.

First stream - CNNs: Four convolutional layers with 16, 32, 64, and 128 filters, respectively, are used in this stream. All convolutional layers use a filter size of 3×3 with same-padding and the ReLU (Rectified Linear Unit) function as the non-linear activation function. Layers 2 and 4 use a stride = 1, while layers 1 and 3 use a stride = 2 to reduce the size of feature maps. The final convolutional layer produces 128 feature maps, each with a dimension of 64×64 . These feature maps are concatenated with the feature maps produced by the LBC layer.

Second stream - LBC layer: The LBC layer is first introduced in [12] to perform as well as traditional convolutional layers but with fewer parameters. The LBC layer also uses convolution operations but with several fixed filters, which requires learning fewer parameters. By using fixed filters, the LBC layer can extract more specific features from a spatial region, which can be merged in an adaptive manner by learning a limited number of parameters. Moreover, using fixed filters allows the LBC layer to reduce memory requirements and thus energy costs, while obtaining a similar level of accuracy to that of traditional convolutional layers [12].

The LBC layer in this work contains two main parts: (1) a set of eight 3×3 filters with non-trainable (fixed) weights, and (2) a set of trainable parameters. The LBC layer first convolves the image with the eight 3×3 fixed filters to produce eight difference-maps after applying the non-linear activation function Sigmoid to map values to the range $[0, 1]$. The eight 3×3 fixed filters effectively extract LBPs as a set of convolution operations with a stride = 1. These filters have values $\in \{-1, 0, 1\}$. Specifically, the value of the central

location of each 3×3 filter is always -1 . For each 3×3 filter, only one location next to the central location has a value of 1, while the remaining seven locations have values of 0. The location with a value equal to 1 changes in each fixed filter, with eight possible locations. Each of the eight difference-maps produced after convolution provides information about how the central location of a 3×3 region of the image compares to the neighbouring location with a value of 1 in the filter (in terms of their pixel values).

The LBC layer linearly merges the eight difference-maps in different ways to produce a single feature map. Our LBC layer uses 32 different ways of merging these eight difference-maps to create a total of 32 feature maps of size 64×64 . Hence, there is a set of 8×32 learnable parameters in our LBC layer. Fig. 2 depicts the complete architecture of our LBC layer. The n^{th} feature map produced by the LBC is then:

$$\mathbf{V}^n = \sum_{m=1}^8 \sigma(\mathbf{b}_m * \mathbf{x}) * \mathbf{w}^n, \quad (1)$$

where \mathbf{b}_m is the m^{th} 3×3 fixed filter, \mathbf{x} is the input image, $*$ denotes a convolution operation, $\sigma(\cdot)$ is the Sigmoid activation function, and \mathbf{w}^n is a vector of eight learnable parameters for the n^{th} feature map.

The computation of the n^{th} feature map, \mathbf{V}^n , is implemented as a convolution operation with filter with size $1 \times 1 \times 1$ and a stride = 4 across the eight difference-maps.

FC layers: We concatenate the feature maps produced by the first stream (CNNs) with the feature maps produced by the second stream (LBC layer) to give a total of 160 feature maps of size 64×64 . These are used as input to two FC layers with 512 and 256 neurons, respectively. The output layer has six neurons to represent the value of the six basic human emotions. In the output layer, Softmax is used as the activation function to represent the output as a probability distribution.

To train the framework and find the set of parameters θ that produce a distribution $\hat{D}_i = p(Y | x_i; \theta)$ very similar to D_i , we use the KL divergence loss function:

$$\downarrow KL = \sum_i D_i \ln \frac{D_i}{\hat{D}_i}, \quad (2)$$

where (\downarrow) indicates that the smaller the value the better. In other words, this loss function aims to minimize the distance between the ground truth distribution and the predicted distribution, for all training examples.

IV. PERFORMANCE EVALUATION

We evaluate the proposed EDL-LBCNN framework for the task of predicting the distribution of emotions from facial expression images. We measure the performance in terms of the distance, or similarity, between the predicted and ground truth distributions. To this end, we use KL divergence and Cosine Similarity (CS). The latter is defined as follows:

$$\uparrow CS = \frac{1}{z} \left(\frac{\sum_i^z D_i \hat{D}_i}{\sqrt{\sum_i^z D_i^2} \sqrt{\sum_i^z \hat{D}_i^2}} \right), \quad (3)$$

where z is the total number of test images, D_i is the ground truth distribution for image i , \hat{D}_i is the corresponding predicted distribution, and (\uparrow) indicates that the larger the value the better.

We use the s-JAFFE dataset for performance evaluation, which is a version of the JAFFE dataset [24] but labeled with vectors of real numbers instead of single classes. It is important to note that there are more recent and larger datasets of facial expression images, such as FER [25] and CK+ [26]. However, these datasets are not labeled with a score for each of the emotions present in the images and are thus not suitable for our problem. The s-JAFFE dataset contains 213 face images of 10 women with seven facial expressions. The size of the images is 256×256 . All images are greyscale and taken from a frontal-view. Each image has been scored by 60 individuals to indicate the degree of six basic emotions (happiness, sadness, fear, anger, disgust, and surprise) using a scale from 1 to 5 (5- high, 1-low). The average value of the 60 values assigned to each emotion is used as the final score for each emotion in each image and to compute a distribution of values.

We increase the size of the s-JAFFE dataset by data augmentation. We specifically apply shifting (to the left and right), rotation and horizontally-flipping operations to each image. After data augmentation, the size of the dataset is 2332 images. We randomly take 20% of this dataset to test the framework, while the remaining 80% is used for training. None the training images are present in the test set.

The dataset is pre-processed by first locating facial landmarks using the method in [27], which is based on a cascade regression algorithm and has been shown to be one of the best methods to extract facial landmarks [28]. Based on the detected landmarks, we crop the images to exclude the background. This process allows the framework to extract features only from the facial region.

For all experiments we use the ADAM optimizer with a fixed learning rate = 0.00001. We set the batch size to 50 and train for 200 epochs. We test four different deep learning architectures. The first architecture is a 4-Layer CNN (which is equivalent to the first stream of the EDL-LBCNN framework) with the same FC layers as those used in our proposed framework. This architecture allows examining the performance of a traditional CNN. For the second architecture, we add low-level facial features to the features extracted by the 4-layer CNN of Architecture 1. These low-level facial features are facial landmarks. This second architecture resembles the method in [23], which also uses CNN features and facial landmarks. For the third architecture, we add one LBC layer to Architecture 2 to combine the feature maps computed by the LBC layer with those computed by the CNNs and the facial landmarks. Architecture 4 is our proposed EDL-LBCNN framework.

Table I shows the performance of all evaluated architectures. These results confirm that the LBC layer, which has fewer parameters than the traditional convolutional layer, can produce feature maps that enhance those produced by traditional CNNs. Note that Architecture 3, although based on a greater number

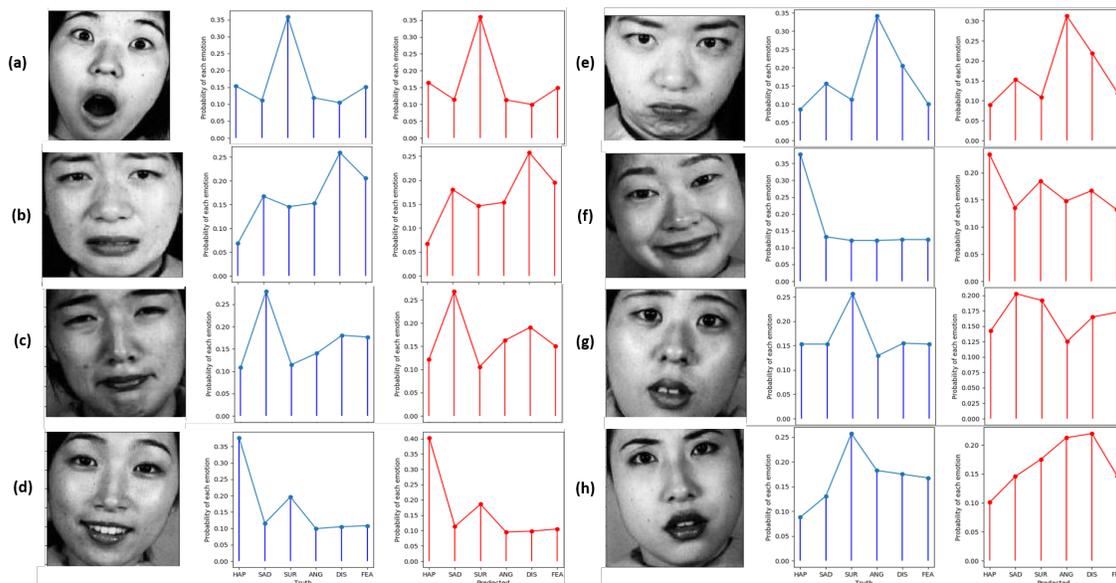


Fig. 3. Some example distributions predicted by our EDL-LBCNN framework. The blue distributions are the ground-truth, while the red distributions are the predicted outputs. HAP, SAD, SUR, ANG, DS, and FEA represent the six basic emotions: Happiness, Anger, Sadness, Surprised, Disgust, and Fear, respectively.

TABLE I
EVALUATION RESULTS OF SEVERAL DEEP LEARNING ARCHITECTURES FOR LDL ON THE S-JAFFE DATASET.

Architecture	Details of the architecture	$\downarrow KL$	$\uparrow CS$
Architecture 1	4-Layer CNN	0.0066	0.9936
Architecture 2	4-Layer CNN + Landmarks	0.0065	0.9937
Architecture 3	EDL-LBCNN + Landmarks	0.0093	0.9915
Architecture 4	EDL-LBCNN	0.0054	0.9949

TABLE II
EVALUATION RESULTS OF SEVERAL METHODS FOR LDL ON THE S-JAFFE DATASET. RESULTS ARE TABULATED AS (MEAN \pm STD.)

Method	$\downarrow KL$	$\uparrow CS$
EDL [10]	0.0745 \pm 0.008	0.9297 \pm 0.007
EDL-LRL [10]	0.0361 \pm 0.004	0.9660 \pm 0.004
EDL-LBCNN	0.0168 \pm 0.003	0.9842 \pm 0.003

of features, does not outperform our proposed framework.

Fig. 3 illustrates some of the results produced by the proposed EDL-LBCNN framework. There are some cases, e.g., (g) and (h), for which our framework does not perform very well. This may be due to the fact that in these cases the expression resembles a neutral expression and the distribution of emotions is hence more uniform. For cases (a), (b), and (d), where a strong and clear facial expression is depicted, our framework performs very well.

We also compare our EDL-LBCNN framework against state-of-the-art LDL methods, EDL [9] and EDL-LRL [10]. Since no data augmentation is used to compute the results reported in [9] and [10], we do not use data augmentation in our framework for this evaluation. Table II tabulates the

TABLE III
EVALUATION RESULTS OF SEVERAL DEEP LEARNING ARCHITECTURES FOR LDL ON THE SBU-3DFE DATASET FOR PREDICTING THE DEGREE OF ATTRACTIVENESS.

Architecture	Details of the architecture	$\downarrow KL$	$\uparrow CS$
Architecture 1	4-Layer CNN	0.0592	0.9717
Architecture 2	4-Layer CNN + Landmarks	0.0511	0.9761
Architecture 3	EDL-LBCNN + Landmarks	0.0514	0.9782
Architecture 4	EDL-LBCNN	0.0500	0.9767

performance of these methods. These results are presented as the *mean \pm std* (standard deviation) over 10 evaluations on the s-JAFFE dataset. The results of EDL and EDL-LRL are reported here as reported in [10]. The EDL-LBCNN framework outperforms these state-of-the-art methods.

Finally, to further confirm the robustness of our framework, we evaluate its performance on a different and yet similar LDL problem: estimating the degree of attractiveness from face images on the SCUT-FBP dataset [29]. This dataset has 500 face images. Each image is scored by 70 individuals using a scale from 1-5 [5-high (very attractive), 1-low (not attractive)], which results in a score distribution. We use the same augmentation methods as those used for the s-JAFFE dataset to increase the number of samples to 4,500 images. As tabulated in Table III, the EDL-LBCNN framework can also learn to estimate a distribution of values very accurately for this problem compared to other architectures. Note that, however, in this case Architecture 3 outperforms the EDL-LBCNN framework in terms of the CS metric. This shows that facial landmarks are useful to estimate the degree of attractiveness, as attractiveness is usually associated with the geometry and symmetry of the face. Although Architecture 3

attains the highest CS value, this architecture uses the LBC layer, which shows that the features extracted by this layer are also strong for this task.

V. CONCLUSIONS

In this work, we have proposed a framework to predict the distribution of basic human emotions from facial expressions depicted in images. We adopted LDL in a deep learning framework to build a general and robust model that uses CNN features. Our proposed framework, called EDL-LBCNN, also incorporates an LBC layer to emphasize texture information, since this information has been shown to be important in recognition tasks involving face images. A series of experiments on the s-JAFFE dataset have shown that the EDL-LBCNN framework is superior to state-of-the-art LDL methods for human emotion recognition.

ACKNOWLEDGMENT

This work has been sponsored by Taif University and The Ministry of Education of Saudi Arabia.

REFERENCES

- [1] E. Granger W. C. D. Melo and A. Hadid. Depression detection based on deep distribution learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2019.
- [2] D. J. Tubbs and K. A. Rahman. Facial expression analysis as a means for additional biometric security in recognition systems. In *Communications in Computer and Information Science Multimedia Communications*, volume Services and Security, pages 113–123, 2015.
- [3] C. C. Ferrer E. Barsoum, C. Zhang and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, 2016.
- [4] S. Li and W. Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. In *International Journal of Computer Vision*, volume 127, pages 884–906, 2018.
- [5] M. Khaled A. T. Ahmed B. R. Ilyas, B. Mohammed and A. Ihsen. Facial expression recognition based on dwt feature for deep cnn. In *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2019.
- [6] D. Chan A. Mollahosseini 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)*, 2016.
- [7] B. Song W. Liu Y. Li N. Zeng, H. Zhang and A. M. Dobaie. Facial expression recognition via learning deep sparse autoencoders. In *Neurocomputing*, volume 273, page 643–649, 2018.
- [8] S. Minaee and A. Abdolrashidi. Deep-emotion: Facialexpression recognition using attentional convolutional network. In *CoRR*, volume abs/1902.01019, 2019.
- [9] H. Xue Y. Zhou and X. Geng. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM international conference on Multimedia*, volume MM '15, pages 1247–1250, New York, NY, 2015.
- [10] W. Li T. Ren, X. Jia and S. Zhao. Label distribution learning with label correlations via low-rank approximation. In *the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.
- [11] J. Sang G. Zheng and C. Xu. Understanding deep learning generalization by maximum entropy. In *IEEE*, 2017.
- [12] V. N. Boddeti F. Juefei-Xu and M. Savvides. Local binary convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR) IEEE*, pages 19–28, 2017.
- [13] M. Pietikäinen T. Ojala and D. Harwood. A comparative study of texture measures with classification based on featured distributions. In *Pattern Recognition*, volume 29(1), pages 51–59, 1996.
- [14] F. Juefei-Xu and M. Savvides. Subspace-based discrete transform encoded local binary patterns representations for robust periocular matching on nist’s face recognition grand challenge. In *IEEE Trans. . . on Image Processing*, volume 23(8), page 3490–3505, 2014.
- [15] X. Geng. Label distribution learning. In *Transactions on Knowledge and Data Engineering IEEE*, volume 28(7), pages 1734–1748, 2016.
- [16] Q. Wang X. Geng and Y. Xia. Facial age estimation by adaptive label distribution learning. In *IEEE International Conference on Pattern Recognition*, page 4465–4470, 2014.
- [17] X. Geng and R. Ji. Label distribution learning. In *IEEE International Conference on Data Mining Workshops*, page 377–383, 2013.
- [18] K. Smith-Miles X. Geng and Z. Zhou. Facial age estimation by learning from label distributions. In *AAAI Conference on Artificial Intelligence*, page 451–456, 2010.
- [19] C. Yins X. Geng and Z. Zhou. Facial age estimation by learning from label distributions. In *IEEE Transactions on Pattern Analysis & Machine Intelligence*, volume 35(10), page 2401–2412, 2013.
- [20] A. Sadvnik K.-C. Peng, T. Chen and A. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*, 2015.
- [21] M. She J. Yang and D. Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.
- [22] C. Xing Z. Huo X. Wei Y. Zhou J. Wu X. Yang, B.Gao and X. Geng. Deep label distribution learning for apparent age estimation. In *International Conference on Computer Vision Workshop (ICCVW) IEEE*, 2015.
- [23] Y. Fan Z. Guo S. Liu, B. Li and A. Samal. Facial attractiveness computation by label distribution learning with deep cnn and geometric features. In *International Conference on Multimedia and Expo (ICME) IEEE*, 2017.
- [24] M. Kamachi M. J. Lyons, S. Akamatsu and J. Gyoba. Coding facial expressions with gabor wavelets. In *International Conference on Automatic Face and Gesture Recognition IEEE*, volume 3rd IEEE, pages 200–205, 1998.
- [25] P. L. Carrier A. Courville M. Mirza B. Hamner W. Cukierski Y. Tang D. Thaler D.-H. Lee et al. I. J. Goodfellow, D. Erhan. Challenges in representation learning: A report on three machine learning contests. In *Neural Networks*, volume 64, page 59–63, 2015.
- [26] T. Kanade J. Saragih Z. Ambadar P. Lucey, J. F. Cohn 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)*, volume 2010 IEEE Computer Society Conference on. IEEE, page 94–101, 2010.
- [27] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Conference on Computer Vision and Pattern Recognition (CVPR) IEEE*, 2014.
- [28] Y. Wu and Q. Ji. Facial landmark detection: A literature survey. In *International Journal of Computer Vision*, 2018.
- [29] L. Jin J. Xu D. Xie, L. Liang and M. Li. Scut-fbp: A benchmark dataset for facial beauty perception. In *IEEE International Conference on Systems*, volume Man, page 1821–1826, 2015.