Facial Expression Recognition Based on LLENet

Dan Meng[1], Guitao Cao*,[1,2]
[1] School of Computer Scinence and Software Engineering
East China Normal University
Shanghai, China 200062
Corresponding Author Email:
*gtcao@sei.ecnu.edu.cn

Zhihai He[2] IEEE Fellow
[2]Department of Electrical and Computer Engineering
University of Missouri
Columbia, USA MO65211

Wenming Cao+,[3]
[3]College of Information Engineering
Shenzhen University
Shenzhen, China 518060
Co-corresponding Author Email:
+wmcao@szu.edu.cn

***Abstract*—Facial expression recognition plays an important role in lie detection, and computer-aided diagnosis. Many deep learning facial expression feature extraction methods have a great improvement in recognition accuracy and robustness than traditional feature extraction methods. However, most of current deep learning methods need special parameter tuning and ad hoc fine-tuning tricks. This paper proposes a novel feature extraction model called Locally Linear Embedding Network (LLENet) for facial expression recognition. The proposed LLENet first reconstructs image sets for the cropped images. Unlike previous deep convolutional neural networks that initialized convolutional kernels randomly, we learn multi-stage kernels from reconstructed image sets directly in a supervised way. Also, we create an improved LLE to select kernels, from which we can obtain the most representative feature maps. Furthermore, to better measure the contribution of these kernels, a new distance based on kernel Euclidean is proposed. After the procedure of multi-scale feature analysis, feature representations are finally sent into a linear classifier. Experimental results on facial expression datasets (CK+) show that the proposed model can capture most representative features and thus improves previous results.**

***Index Terms*—Face expression recognition, deep learning, kernel distance, locally linear embedding (LLE).**

I. Introduction

For many tasks in pattern recognition, such as hand-written digit recognition, face recognition and object recognition, discovering suitable representations of images is a critical problem. Among these visual tasks, facial expression recognition is considered to be the most challenging and long-standing topic since it is widely used in areas of lie detection, surveillance, identity authentication, access control and human-computer interaction. Classical feature extraction methods for face images include histogram of oriented gradients (HOG), local binary pattern (LBP), Gabor features and their fusions. These low-level feature extraction methods often lack the ability of capturing the most useful information directly from data samples. As a result, although these feature representations work well on certain face recognition tasks, they can hardly cope with uncontrolled environments (e.g. occlusion, large variations in illumination and pose, and accessory changes).

Over the past few years, feature extraction methods based on manifold learning and deep learning have been made breakthrough on a wide range of vision tasks [1], [2], [3]. It is worth notice that, Chan [3] developed a very simple DNN for computer vision classification tasks, called PCANet. PCANet [3] utilized the most basic and easy processing components to emulate DNN, which is able to train and adapt to various tasks easily. However, such a simple architecture ignores the valuable known label information of the sample images during training stage, thus affects the recognition accuracy. Besides, Chan [3] applied linear dimension reduction method (PCA) to find kernels (detectors) with maximum responses, which might be unsuitable for facial expression recognition, given face images' nonlinear manifold structure [1]. Enlightened by PCANet [3], in this paper we introduce a novel feature extraction model called LLENet for face expression recognition to take advantage of features learned through deep learning method while overcome drawbacks of PCANet [3]. The architecture of the proposed feature extraction model is simple, since it only consists of a procedure of image reconstruction sets (IRS), a multi-stage feature convolutional layer and a multi-scale feature analysis layer. Besides, we remove back propagation from standard DNN, so we do not need special parameter tuning and ad hoc fine-tuning tricks.

The rest of the paper is organized as follows. Section II presents the proposed LLENet, and experimental results are provided in Section III followed by conclusions in Section IV.

II. Our Method (LLENet)

Given face video sequences or images, we use a facial detector to localize the face region of each frame or image. The localized face region is then cropped and normalized to a fixed height and width, generating a face image. To extract feature representations, face images in IRS are convoluted with learned kernels before multi-scale feature analysis. These feature representations are then fed into classification model. Figure 1 summarizes our proposed LLENet.

A. Image Reconstruction Sets

Let $\{(X_i, L_i)\,|i = 1, 2, \cdots, N_{train}\}$ be training images in the dataset, where $X_i \in R^{m\times n}$ represents the i-th training sample, and $L_i = 1, 2, \cdots, C$ is the class label of X_i. For each training sample X_i, we take a $k_1 \times k_2$ patch with a step of 1 pixel around each pixel, and collect overlapped patches of i-th image $\tilde{X}_i = \left[x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,\tilde{m}\tilde{n})}\right]$, where $x_{(i,j)}$

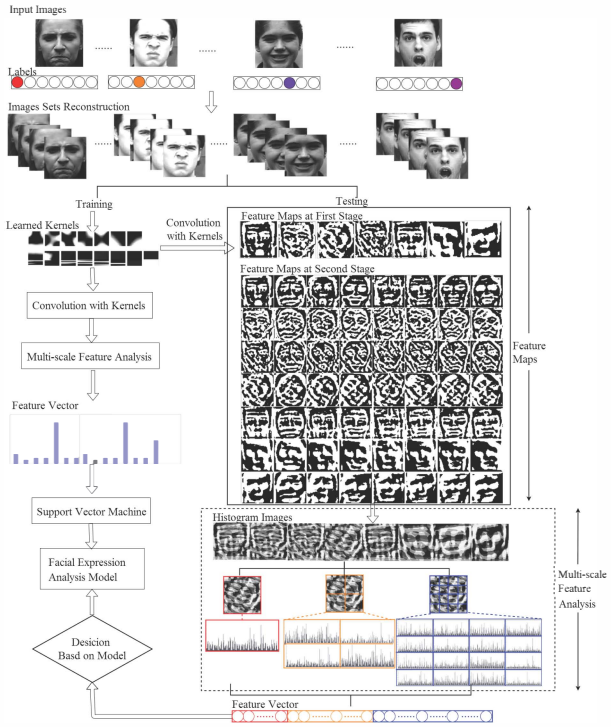

Fig. 1. Flowchart of face recognition framework based on LLENet.

denotes the j-th vectorized patch in X_i, $\tilde{m} = m - \lceil k_1/2 \rceil$, $\tilde{n} = n - \lceil k_2/2 \rceil$, and $\lceil z \rceil$ gives the smallest integer greater than or equal to z. Let $\tilde{X}_i^{(p)} \in R^{1 \times \tilde{m}\tilde{n}}$ denotes p-th row in $\tilde{X}_i$, which can be seen as a copy of the original image with some rows and columns removed from X_i when we reshape vector $\tilde{X}_i^{(p)}$ to a $\tilde{m} \times \tilde{n}$ matrix. We call $\tilde{X}_i = \left[\tilde{X}_i^{(1)}; \tilde{X}_i^{(2)}; \cdots; \tilde{X}_i^{(k_1 k_2)}\right]$ a reconstructed image set, and the corresponding labels can be represented as $\tilde{L}_i = \left\{\tilde{L}_i^{(p)} = L_i, p \in [1, k_1 k_2]\right\}$. By constructing the same image set for all training samples, we finish reconstructing image sets $\tilde{X} = \left\{\tilde{X}_1; \tilde{X}_2; \cdots; \tilde{X}_{N_{train}}\right\}$. The labels of the reconstructed image sets are denoted by $\tilde{L} = \left\{\tilde{L}_1; \tilde{L}_2; \cdots; \tilde{L}_{N_{train}}\right\}$.

B. Improved LLE Based on the Modified Kernel Distance

Among all the image distance metrics d used in LLE algorithm, Euclidean distance d_E or its kernel form d_{Ek} are the most commonly used due to their simplicity. However, this distance measurement suffers from a high sensitivity even to small deformation. This paper proposes a modified kernel distance, which is called Kernel Nonlinear Distance d_{KND}. The modified distance d_{KND} is defined as:

$$d_{KND} = \left(\exp^{\frac{d_{Ek}^2}{mean(d_{Ek})}}\right)^{\frac{1}{2}} + \alpha \max\left(d_{Ek}\right). \quad (1)$$

where $d_{Ek}\left(K_{IiCDM}(i,j)\right) = \sqrt{\left\langle \phi\left(K_{IiCDM}^i\right) - \phi\left(K_{IiCDM}^j\right), \phi\left(K_{IiCDM}^i\right) - \phi\left(K_{IiCDM}^j\right)\right\rangle}$ denotes kernel Euclidean distance, and we use classical

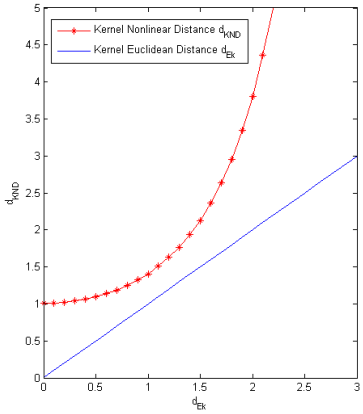

Fig. 2. Comparison of distance curve between d_{KND} and d_{Ek}, where $k_1 k_2 = 256$, and $\alpha = 1/256$.

Guass kernel function as the kernel function. Besides, $K_{IiCDM} = \frac{S_{inter}}{S_{intra}}$, where S_{intra} and S_{inter} are the intra and inter scatter matrix for IRS images respectively. In Equation 1, $mean\left(d_{Ek}\right)$ and $\max\left(d_{Ek}\right)$ are the average and maximum kernel Euclidean distance between all pairs of data points respectively.

In order to have a better understanding of the advantage of the proposed nonlinear distance, Figure 2 shows a comparison of distance curves between d_{KND} and d_{Ek} when the original kernel Euclidean distance d_{Ek} increased linearly in the range of $[0, 3]$. From Figure 2, we can see that the modified distance d_{KND} brings distance between different convolutional kernels into an exponential growth mode. The reconstruction error of the original LLE can be rewritten as follows:

$$\varepsilon\left(W_i\right) = W_i^T d_{KND_i} W_i. \quad (2)$$

Thus objective cost function can be written as

$$\min J\left(V\right) = tr\left(V^T \Phi\left(K_{IiCDM}\right) M \Phi\left(K_{IiCDM}\right)^T V\right). \quad (3)$$

The solution is known as the eigenvectors corresponding to the smallest d eigenvalues of M. Once V is obtained, for test samples, the nonlinear feature maps are given as $\tilde{Y}_i = \tilde{X}_i * V^T$.

C. Multi-scale features analysis

For each set of feature maps, we convert the feature maps belonging to the corresponding filter for the last stage into one histogram image whose every pixel is an integer in the range $[0, 255]$, and treated as a distinct "word". For each set of B^τ output images X_{il}^τ belongs to the V_l^τ at final stage τ, we can obtain by:

$$X_{il}^\tau = X_i^{\tau-1} * mat_{k_1 k_2}\left(V_l^\tau\right)^T \quad (4)$$

where $i \in \left[1, N_{train} \prod_{t=1}^{\tau-1} B^t\right], l \in [1, B^\tau]$, V_l^τ is the l-th filter at final stage τ, and $mat_{k_1 k_2}\left(V_l^\tau\right)$ is a function that

maps $V_l^\tau \in R^{k_1 k_2}$ to a maxtrix of dimension of $R^{k_1 \times k_2}$. Then histogram image $Hist_l^\tau$ can be expressed as:

$$Hist_{il} = \sum_{i=1}^{\prod_{t=1}^{\tau-1} B^t} \left(2^{\tau-1} \mod 256\right) \times H\left(X_{il}^\tau\right). \quad (5)$$

where $H\left(X_{il}\right)$ is a heviside step function. For every B^τ histogram image $Hist_{il}$, we construct a sequence of grids at resolutions $0, 1, \cdots, Q$, such that the grid at level q has 2^q cells along each dimension, with a total of $G = \sum_{q=0}^{Q} 2^q$ cells. Let $Ch_g^q\left(Hist_{il}\right)$ denotes the vector containing numbers of points from $Ch_g^q\left(Hist_{il}\right)$ that fall into g-th cell at resolution q according to different "words". We cascade $Ch_g^q\left(Hist_{il}\right)$ to build a multi-scale feature map for histogram image as:

$$Ch\left(Hist_{il}\right) = \left[Ch_1^0\left(Hist_{il}\right), Ch_2^1\left(Hist_{il}\right), \cdots, Ch_G^Q\left(Hist_{il}\right)\right]. \quad (6)$$

where $Ch\left(Hist_{il}\right) \in R^{G \times 2^8}$. So feature representations for X_i can be expressed as:

$$f_{train}^i \doteq \left[Ch\left(Hist_{i1}\right), Ch\left(Hist_{i2}\right), \cdots, Ch\left(Hist_{iB^\tau}\right)\right]. \quad (7)$$

III. Experiment

To verify the proposed method, we performe experiments on CK+ to evaluate the performance of the proposed algorithm using linear SVM for classification. The CK+ [4] dataset consists of 593 sequences from 123 subjects of both sexes. We cropped facial expression images for each sequence by a face detector. Then we constructed three image pairs as done in [2] for each sequence. The first image in each pair corresponds to the onset image (neutral) and the other two images correspond to the two peak expression faces. As for the output facial expression images, the peak expression faces in image pairs were then normalized to 128×128 pixels, with labels corresponding to the peak expression faces.

We randomly select 70% samples to learn the convolutional kernels, and the rest samples extract features based on the learned kernels. Table I shows the comparison of the proposed LLENet and state-of-the-art methods. From Table I we can see that our proposed LLENet outperforms all the other methods and achieves better performance than the least STRBM [2]. Additionally, we present the confusion matrix of our method in Figure 3. From Figure, one can observe that LLENet achieves best performance on happy, disgust, and contempt, and worst performance on fear.

IV. Conclusion

In this paper, we proposed a novel feature extraction model based on supervised kernel learning and improved LLE for facial expression recognition. The proposed feature extraction model has three major components: image sets reconstruction (IRS), multi-stage supervised kernel learning, and multi-scale feature analysis. Experimental results show that the proposed feature extraction model is able to obtain robust and efficient feature representations for facial expression recognition task.

TABLE I
Comparison with state-of-art on CK+ dataset.

Method	Recognition Rate(%)
SPTS+CAPP [4]	83.30
CLM-SRI [5]	88.60
EAI [6]	82.60
LDN [7]	89.30
STM-ExpLet [8]	94.19
STRBM [2]	95.66
PCANet [3]	94.31
LLENet	**96.94**

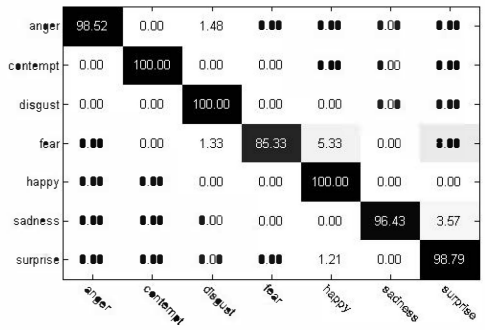

Fig. 3. Confusion Matrix of LLENet-2 on CK+ dataset.

Acknowledgment

This work was supported by Natural Science Foundation of China (61375015).

References

[1] G. Q. Wang, N. F. Shi, and Y. X. Shu, "Embedded manifold-based kernel fisher discriminant analysis for face recognition," *Neural Process Letters*, vol. 43, no. 1, pp. 1–16, 2016.

[2] S. Elaiwat, M. Bennamoun, and F. Boussaid, "A spatio-temporal RBM-based model for facial expression recognition," *Pattern Recognition*, vol. 49, pp. 152–161, 2016.

[3] T. H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 12, pp. 5017–5032, 2015.

[4] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "Comprehensive data base (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, Jun. 2010, pp. 94–101.

[5] L. Jeni, D. Takacs, and A. Lorincz, "High quality facial expression recognition in video streams using shape related information only," in *IEEE International Conference on Computer Vision (ICCV) Workshops*, Nov. 2011, pp. 2168–2174.

[6] S. Yang and B. Bhanu, "Understanding discrete facial expressions in video using an emotion avatar image," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 4, pp. 980–992, Aug. 2012.

[7] A. R. Rivera, J. R. Castillo, and O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Transactions on Image Processing (TIP)*, vol. 22, no. 5, pp. 1740–1752, 2013.

[8] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expression let son spatio-temporal manifold for dynamic facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.