2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Hand-crafted Feature Guided Deep Learning for Facial Expression Recognition

Guohang Zeng, Jiancan Zhou, Xi Jia, Weicheng Xie and Linlin Shen
School of Computer Science & Software Engineering, Shenzhen University, China
ghzeng.cs@hotmail.com, {2150230421,jiaxi}@email.szu.edu.cn, {wcxie,llshen}@szu.edu.cn

***Abstract*— A number of facial expression recognition algorithms based on hand-crafted features and deep neutral networks have been developed. Motivated by the similarity between the hand-crafted features and features learned by deep network, a new feature loss is proposed to embed the information of hand-crafted features into the training process of network, which tries to reduce the difference between the two features. Based on the feature loss, a general framework for embedding the traditional feature information was developed and tested using CK+, JAFFE and FER2013 datasets. Experimental results show that the proposed network achieves much better accuracy than the original hand-crafted feature and the network without using our feature loss. When compared with other algorithms in literature, our network also achieved the best performance on CK+ dataset, i.e. 97.35% accuracy has been achieved.**

I. INTRODUCTION

While hand-crafted features use filters like Gabor, Local Binary Pattern (LBP) and SIFT (Scale Invariant Feature Transform) to extract features from images, deep neural network (DNN) use end-to-end models to automatically learn filters at different levels for feature extraction. As the optimization space of DNN can be very large due to the large number of network parameters, some algorithms employ local searching strategies to speed up the searching efficiency.

Tudor Ionescu et al. [1] proposed a local feature learning approach using neighbors around each testing sample for facial expression recognition. Jung et al. [2] introduced a fine-tuning network for expression recognition by integrating the geometry features into texture feature network and fine-tuned the last two layers of the network. Sikka et al. [3] used locality-constrained linear coding and max-pooling strategy to extract multi-scale dense SIFT feature for expression recognition. Zadeh et al. [4] constructed a local model with convolutional experts constraint by incorporating a set of appearance prototypes of different poses and expressions for landmark detection. Pan et al. [5] used orthogonal projection layer to replace a pooling layer by adding an orthogonal projection constraint on the loss function, so as to reduce the model size and redundancy of convoluted features. These algorithms employed a fine-tuning strategy or imposed a constraint on the deep network without using the merits of hand-crafted features. Fusing the deep feature with hand-crafted features is another approach to improve the performance of both features.

Qian et al. [6] proposed to fuse the higher-layer feature of a deep network and the hand-crafted features like PHOG and LBP-EOH for vehicle classification. Liu et al. [7] concatenated hand-crafted HOG and dense SIFT features with deep CNN feature for expression recognition. Paul et al. [8] merged the deep features of the top five layers and hand-crafted quantitative features for survival prediction. Suggu et al. [9] concatenated several hand-crafted features with the convolution output before the fully connected (FC) layer for the network training, while Majtner et al. [10] fused the discrimination probabilities of the hand-crafted and deep features after SVM for final recognition.

However, these algorithms directly concatenated two categories of features without embedding the information of hand-crafted features into the deep network [11]. Actually, Khorrami et al. [12] showed that the learned deep features are analogous to the basic facial action units of expression faces, and Zeiler and Fergus [13] revealed that the learned features in the shallow layers of CNN are similar to hand-crafted features such as Gabor feature [14]. Juefei-Xu et al. [15] replaced convolution layer with LBP like operator and largely reduced the model complexity with similar performance. More precisely, the hand-crafted and deep features may be similar and complementary, which motivates us to boost the performance of deep network by local searching around the hand-crafted feature. Meanwhile, a few deep metric learning algorithms have been proposed to embed constraint information into the loss function of deep networks to improve the performance, such as the CenterLoss [16], SphereFace loss [17], adaptive deep metric learning [18].

For the application of facial expression recognition, many hand-crafted and deep features have been proposed in the recent decades. Examples are maximum margin projection [19], radial feature [20], and deep features such as AU deep network [21], deep neural network (DNN) [22], etc. However, works fusing the hand-crafted and deep features for expression recognition have not been widely studied.

In this work, a new deep network based on hand-crafted feature guidance is proposed for the expression recognition, which consists of the HoloNet network with feature loss (HNwFL) for feature learning and fusion network for recognition. In HNwFL, the hand-crafted feature information is embedded into the Hololet network by imposing a new loss metric into the loss function. The motivation of the new loss is to use a hand-crafted feature to guide the network learning

0-7695-6381-3/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00068

and reduce the optimization space during the early training, and add human prior knowledge into the learning of DNN features. In fusion network, the learned feature and the hand-crafted feature are input into a fusion network for the final recognition.

This work makes the following contributions.

- A novel loss metric, namely feature loss, is proposed to provide complementary information for the deep feature during network early training;
- A deep metric learning based framework is proposed to embed the guidance of hand-crafted features into deep networks to improve recognition performance;
- The proposed algorithm achieves competitive performance on two public expression databases compared with the hand-crafted approach, the deep network without the proposed loss and the state-of-the-art algorithms.

This paper is structured as follows. Section II gives a description about the proposed algorithm step by step. The experimental results of the proposed algorithm on public databases are presented in Section III. Finally, discussions and conclusions are addressed in Section IV.

II. THE PROPOSED ALGORITHM

In this section, the proposed feature loss, the deep network structure, the network training and fusion network for recognition are introduced.

A. The Feature Loss

In the proposed network, besides of SoftMax loss and the Center loss [16], a new metric, namely feature loss is proposed to measure the network performance, which is presented in Fig. 1. To benefit the presentation of these losses, the entire network in Fig. 1 is divided into two sub-networks, abbreviated as **MI** and **MII**, whose loss criterions are recorded with **CI** and **CII**, respectively.

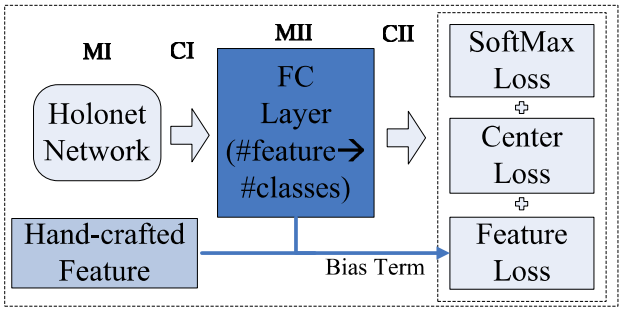

Fig. 1: The proposed network based on SoftMax, center and feature losses. #feature and #classes denote the feature dimension and the number of expression classes, respectively.

Different from the other two losses, the proposed feature loss embeds the information of hand-crafted features for local searching. To study their differences, these three loss

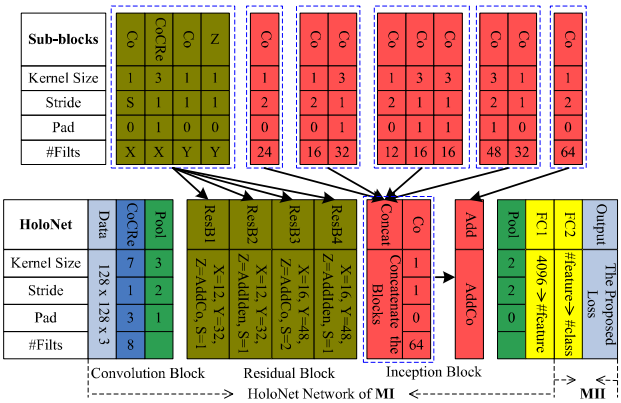

Fig. 2: HoloNet network [26] with adaptive fully connected layer. Co, FC denote the convolution and FC layers. $CoCRe$ denotes the convolution followed by the spatial batch normalization and CReLU. $AddCo$ or $AddIden$ denotes the convolution or identity mapping used in the i-th residual block ($ResBi$).

functions are formulated as follows

$$\begin{cases} \mathcal{L}_S = -\sum_i log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_j e^{W_j^T x_i + b_j}}, \\ \mathcal{L}_C = \frac{1}{2}\sum_i ||x_i - c_{y_i}||, \\ \mathcal{L}_F = \frac{1}{2}\sum_i ||x_i - ||x_i||_2 \cdot g_i||. \end{cases} \quad (1)$$

where $\mathcal{L}_S$, $\mathcal{L}_C$, $\mathcal{L}_F$ are the SoftMax, center and feature guided losses, respectively; W, b are the weight matrix and the bias terms of **MII**; x_i is the **MI** output of the i-th sample and y_i is its expression label, c_{y_i} is the center vector of the **MI** outputs with respect to (w.r.t.) the y_i-th class, g_i is the normalization of the hand-crafted feature f_i of the i-th sample, and can be formulated as follows

$$g_i = \frac{f_i}{||f_i||_2}. \quad (2)$$

Since Gabor features have been widely used in expression recognition and achieved good performance, the hand-crafted feature of Gabor Surface Feature (GSF) [23], [24] is employed in this work. More precisely, the feature of the i-th expression sample is first represented as

$$fo_i = (pf_{p_1}, \cdots, pf_{p_n}), \quad (3)$$

where n is the number of face patches, pf_{p_j} is the GSF representation of the j-th patch. Each patch feature pf_{p_k} is further reduced to six dimensions using principal components analysis (PCA) and linear discriminant analysis (LDA) [25]. Consequently, the hand-crafted feature of the i-th sample is presented as follows

$$f_i = P_{LDA}(P_{PCA}(fo_i)). \quad (4)$$

where P_{PCA}, P_{LDA} are the projection matrixes of PCA and LDA, respectively. For boosting the GSF feature extraction, randomly selected 30 augmented expression samples from the training dataset for each person ID are used for the

projection matrix computation.

B. Network Structure

The proposed feature loss $\mathcal{L}_F$ is further employed in the HoloNet network [26], thereafter abbreviated as HNwFL, i.e. HoloNet network with feature loss, for expression recognition. HoloNet [26] achieves the best recognition rate on the 2016 EmotiW challenge [27], which mainly contains three blocks, i.e. the phase-convolution, the phase-residual and the inception-residual blocks. In the proposed algorithm, the HoloNet (the last FC layer was removed) is used as network **MI** in Fig. 1, and is presented in Fig. 2.

In HoloNet, the phase-convolution block with modified Concatenated Rectified Linear Unit (CReLU) is proposed to maintain the positive and negative phase information after convolution. The phase-residual blocks are proposed to obtain a relatively high accuracy for high depth network structure. The inception-residual block with multiple sizes of convolution kernels is proposed to construct multi-scale features.

In order to compute the proposed feature loss $\mathcal{L}_F$ in equation (1), the number of output neurons in the last FC layer of **MI** is set the same as the dimension of the hand-crafted feature f_i in equation (4).

C. Network Training

With the proposed feature loss, the final loss of the proposed network is formulated as follows

$$\min \mathcal{L} = \mathcal{L}_S + \lambda_C \mathcal{L}_C + \lambda_F \mathcal{L}_F. \tag{5}$$

where λ_C, λ_F are the regularization parameters, which are decreased with fixed decaying factors as the network training proceeds.

The minimization of the loss function in equation (5) is formulated as a fitting form. To backward the entire network optimization, the forwarding and backwarding operations of each model (**MI** or **MII**) and criterion (**CI** or **CII**) are used. While only SoftMax loss is considered for **CII**, all the losses are included for **CI**. The model forwarding gives the output for each layer; then the criterion forwarding computes the final loss function based on the network output; the criterion backwarding obtains the derivatives of the loss function w.r.t. the network output; finally, the model backwarding computes the derivatives of the loss function w.r.t. the network input and the network weight parameters.

For the network optimization, the gradient of $\mathcal{L}$ w.r.t. each variable is obtained as follows

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}_S}{\partial W}, \\ \begin{cases} \frac{\partial \mathcal{L}}{\partial x_i} = \frac{\partial \mathcal{L}_S}{\partial x_i} + \lambda_C \frac{\partial \mathcal{L}_C}{\partial x_i} + \lambda_F \frac{\partial \mathcal{L}_F}{\partial x_i}, \\ \frac{\partial \mathcal{L}_C}{\partial x_i} = x_i - c_{y_i}, \\ \frac{\partial \mathcal{L}_F}{\partial x_i} = x_i - ||x_i||_2 \cdot g_i, \end{cases} \\ \begin{cases} \frac{\partial \mathcal{L}}{\partial c_j} = \lambda_C \sum_i \gamma_{i,j} \frac{\partial \mathcal{L}_C}{\partial c_j}, \\ \frac{\partial \mathcal{L}_C}{\partial c_j} = c_j - x_i, \\ \gamma_{i,j} = \frac{\delta(y_i = j)}{1 + \sum_k (y_k = j)}, \end{cases} \\ \frac{\partial \mathcal{L}}{\partial \varpi} = \Sigma_i \frac{\partial \mathcal{L}}{\partial x_i} \frac{\partial x_i}{\partial \varpi}. \end{cases} \tag{6}$$

where $\gamma_{i,j}$ is the regularization term suggested in the work [16], $\delta(\cdot)$ is the Dirac delta function, and ϖ is the network parameters of **MI**. Since Cross-Entropy function is employed for the SoftMax loss $\mathcal{L}_S$, the partial derivatives $\frac{\partial \mathcal{L}_S}{\partial x_i}$, $\frac{\partial \mathcal{L}_S}{\partial W}$ are automatically obtained by the network backward of **MII**.

With the obtained gradients in equation (6), the network parameters are iteratively updated with optimization of S-tochastic Gradient Descent (SGD) as follows

$$\begin{cases} c_j^{t+1} = c_j^t - \alpha \frac{\partial \mathcal{L}}{\partial c_j}, \\ W^{t+1} = W^t - \mu_2^t \frac{\partial \mathcal{L}}{\partial W^t}, \\ \varpi^{t+1} = \varpi^t - \mu_1^t \frac{\partial \mathcal{L}}{\partial \varpi^t}, \end{cases} \tag{7}$$

where α is the learning rate of the centers $\{c_j\}$, μ_1^t, μ_2^t are the learning rates w.r.t. **MI**, **MII**, respectively. The proposed network is not sensitive to the bias parameters $\{b_j\}$, thus, these parameters are always set to zero for simplification.

For the database with unbalanced proportions, weight balancing strategy is employed, i.e. the weight of each class is inversely proportional to the number of samples as follows

$$w_i = \frac{1/\#S_i}{\sum_i 1/\#S_i}. \tag{8}$$

where $\#S_i$ denotes the number of expression samples of the i-th class. The SoftMax loss function in equation (1) is then formulated as follows

$$\mathcal{L}_S = -\sum_i w_{y_i} log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_j e^{W_j^T x_i + b_j}}. \tag{9}$$

In this way, the recognition rates of all the classes with different sample sizes are balanced.

For small sample data, data augmentation strategy is employed to increase training samples and decrease the influence of face misalignment. The database is augmented by first flipping the faces, and capturing $3 \times 3 \times 3$ different face regions with the corresponding binary masks for each expression image. The final expression label is predicted by probability voting, i.e. the label corresponding to the largest cumulative probability is selected as the predicted expression label.

For clarity, the entire optimization framework of the proposed HNwFL is illustrated in Algorithm 1.

D. Fusion Network for Classification

After feature learning with the proposed network shown in Fig. 1, the concatenation of **MI** output and the hand-crafted feature is further input into a fusion network for recognition, which is shown in Fig. 3. The fusion network and criterion are abbreviated as **MIII** and **CIII**, whose parameter optimization is similar to that of parameter W in the network **MII**. Compared with **MIII**, **MI+MII+MIII** used the information of a hand-crafted feature to guide the network learning; compared with **MI+MII**, **MI+MII+MIII** added a fine tuning network of the fused feature to boost the discrimination ability. The entire optimization framework of the fusion network is illustrated in Algorithm 2.

Algorithm 1 The training of HNwFL.

Set the weights of each class with equation (8).
Obtain GSF features $\{f_i\}$ of all the samples by equation (4).
Set the parameters λ_C=1e-4, λ_F=1e-4, $MaxIter = 2e2$.
Initialize the network parameters ϖ of **MI**, the weight vector W and c_j of **MII**.
for $s = 0, \cdots, MaxIter$ **do**
  Perform **MI** forward to obtain the output x_i.
    Perform **MII** forward to obtain the output z_i.
    Perform **CII** forward to obtain the loss $\mathcal{L}_S$ using z_i.
    Perform **CII** backward to obtain the gradient $\frac{\partial \mathcal{L}_S}{\partial z_i}$.
    Perform **MII** backward to compute the gradient $\frac{\partial \mathcal{L}_S}{\partial x_i}$.
    Perform SGD in equation (7) to renew W.
  Store the L_2 norm of x_i as $||x_i||_2$.
  Perform **CI** forward to obtain the entire loss function $\mathcal{L}$ as in equation (5).
  Perform **CI** backward to obtain $\frac{\partial \mathcal{L}}{\partial x_i}$ as in equation (6) using $||x_i||_2$.
  Perform **MI** backward to compute the gradients $\frac{\partial \mathcal{L}}{\partial \varpi}$ of model **MI**.
  Perform SGD in equation (7) to renew ϖ.
  Renew centers c_j using the gradient $\frac{\partial \mathcal{L}}{\partial c_j}$ in equation (6).
end for
Output the GSF-guided feature $\{x_i\}$ of **MI** output.

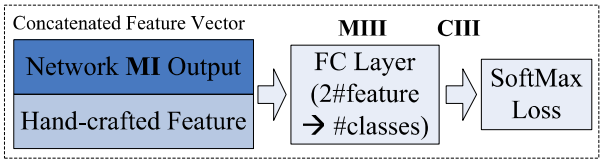

Fig. 3: The fusion network.

III. EXPERIMENTAL RESULTS

We perform the experiments using four-kernel Nvidia TITAN GPU Card, Torch platform and LUA language [28]. The parameter settings of HNwFL and the fusion network are presented in Table I.

The proposed algorithm is tested on the expression databases of the Extended Cohn-Kanade Dataset (CK+) [29], JAFFE [30] and FER2013 database [31], whose examples are presented in Fig. 4. The JAFFE database [30] consists of 213 expression images of 10 Japanese female models, which can be categorized to six basic and the neutral expressions, i.e.

Algorithm 2 The training of the fusion network.

Set the weights of each class with equation (8).
Concatenate GSF features $\{g_i\}$ with **MI** output $\{x_i\}$ to form the input of the network **MIII**.
Initialize the weight vector W of **MIII**, $MaxIter = 1e2$.
for $s = 0, \cdots, MaxIter$ **do**
  Perform **MIII** forward to obtain the output z_i.
  Perform **CIII** forward to obtain the loss $\mathcal{L}_S$ using z_i.
  Perform **CIII** backward to obtain the gradient $\frac{\partial \mathcal{L}_S}{\partial z_i}$.
  Perform **MIII** backward to renew the gradients $\frac{\partial \mathcal{L}_S}{\partial W}$.
  Perform SGD to renew the parameters W.
end for
Output the final recognition labels with the maximum response.

TABLE I: The parameter setting of HNwFL and the fusion network.

Model	Parameter Name	Parameter Value
HNwFL	α	1e-2
	$\mu_1^t = \mu_2^t$	1e-2
	Learning rate	5e-3
	Batch size	100
	Momentum	0.8
	L_2 regularization coefficient	1e-4
	λ_C and λ_F decaying factor	0.8
	Image size	128x128
Fusion Network	Learning rate	1e-1
	Learning decaying factor	1e-3

angry (An), disgust (Di), fear (Fe), happy (Ha), sad (Sa) and surprise (Su). The CK+ database consists of 593 expression sequences from 123 subjects, where 327 sequences are labeled with one of seven expressions (angry, disgust, fear, happy, sad, surprise and contempt). Each sequence contains a set of captured frames when the subject changes his expression, 1033 expression images, i.e., the neutral and three non-neutral images sampled from each expression sequence are used for testing. The FER2013 database [31] consists of 35887 grayscale face images with size 48x48, which are collected from the internet and used for a challenge. The faces were labeled with one of seven categories. The training set consists of 28,709 examples, while the validation and testing sets consist of 3,589 samples individually. Five landmark points were located with [32] for face alignment.

For the following experiment, the person-independent strategy with ten-fold setting is employed for testing and comparison. More precisely, the considered database is divided into ten groups with approximately equal number of

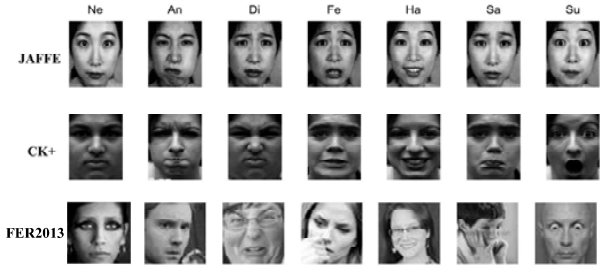

Fig. 4: Examples of three expression databases.

person IDs. While nine of them were used for training, the remaining group was used for testing. The process was randomly repeated ten times and the average accuracy is recorded as the final result. For the batch-based SGD optimization, the training samples are randomly perturbed. For PCA in the GSF construction (4), the feature vectors with the accumulative contribution rate of 0.99 are used.

A. *Performance Evaluation*

To test the performance of the proposed loss function, the network in Fig. 1 is trained with different loss functions, i.e. the SoftMax, the center loss, the proposed feature loss and their different combinations, then the learned features are input into the fusion network for recognition. The recognition rates for different loss functions are presented in Table II. Table II shows that both center and feature losses are beneficial to the recognition rate and the fused loss metric achieves the best performance, i.e. the combination of three losses achieves improvements of 4.82%, 5.64% and 4.16% on CK+, JAFFE and FER2013 databases, respectively, over the SoftMax loss. Meanwhile, the loss with GSF feature guidance performs better than the center loss. Fig. 5 also shows the variation of overall accuracy of the proposed algorithm with different λ_C and λ_F, for JAFFE database.

We also compared the proposed algorithm with GSF and HoloNet with SoftMax loss in Table III, where the ten-fold and mean performance of these algorithms on the CK+ and JAFFE databases are listed. Table III shows that the recognition rate is improved for most folds of both CK+ and JAFFE databases. Meanwhile, the mean recognition rates of the proposed algorithm are 7.25% and 4.85% higher than that of GSF and HoloNet on CK+ database.

To evaluate the overall performance, the confusion matrixes of the proposed algorithm on the databases CK+ and JAFFE are presented in Tables IV and V. Tables IV and V show that the expressions 'angry', 'disgust', 'fear', 'sad' are relatively more difficult than the other three expressions.

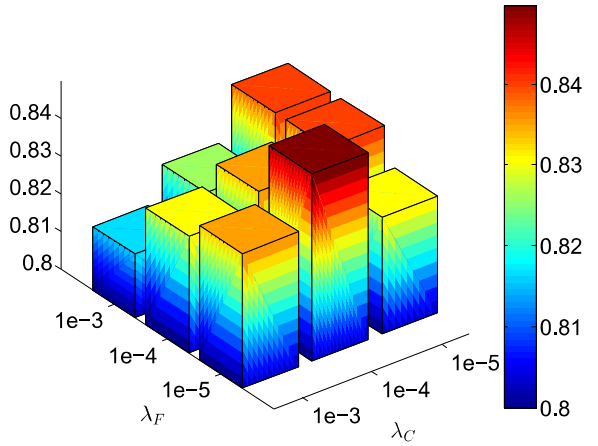

Fig. 5: The average recognition rates of the proposed algorithm on the JAFFE database with different λ_C and λ_F.

For the CK+ database, the expression 'angry' and 'sad' are easy to be confused with the 'neutral' expression; while for the JAFFE database, 'disgust' and 'fear', 'disgust' and 'sad', 'fear' and 'sad' are easy to be confused.

TABLE IV: Confusion matrix (%) of the proposed algorithm on CK+ database.

Exp.	Ne	An	Di	Fe	Ha	Sa	Su
Ne	**98.11**	0.94	0	0	0	0.94	0
An	6.66	**86.66**	4.44	0	0	2.22	0
Di	0	0	**100**	0	0	0	0
Fe	0	0	0	**100**	0	0	0
Ha	0	0	0	0	**100**	0	0
Sa	7.14	0	0	0	0	**92.85**	0
Su	1.20	0	0	0	0	0	**98.79**

TABLE V: Confusion matrix (%) of the proposed algorithm on JAFFE database.

Exp.	Ne	An	Di	Fe	Ha	Sa	Su
Ne	**86.66**	0	0	3.33	3.33	0	6.66
An	0	**93.33**	6.66	0	0	0	0
Di	0	0	**79.31**	10.34	0	10.34	0
Fe	0	0	9.37	**78.12**	0	12.5	0
Ha	9.67	0	0	0	**80.64**	0	9.67
Sa	0	3.22	3.22	9.67	3.22	**80.64**	0
Su	10	0	0	0	3.33	0	**86.66**

To study the effects of HNwFL (**MI+MII**) and the fusion (**MIII**) network on the final expression recognition, different network structures are tested on the three databases and the results are presented in Table VI. When only **MIII** is employed, the GSF is set as the input of the network for recognition. It can be seen from Table VI that both HNwFL

TABLE II: The performance of different losses.

Database	Parameter setting and the average recognition rates (%).			
	λ_C=0, λ_F=0 SoftMax	λ_C=1e-4, λ_F=0 SoftMax+Center Loss	λ_C=0, λ_F=1e-4 SoftMax+Feature Loss	λ_C=1e-4, λ_F=1e-4 SoftMax+ Center Loss+ Feature Loss
CK+	92.53	93.73	94.94	**97.35**
JAFFE	77.93	79.81	82.16	**83.57**
FER2013	57.70	59.04	60.66	**61.86**

TABLE III: The ten-fold and mean recognition rates of GSF, HoloNet with SoftMax and the proposed algorithm on the databases of CK+ and JAFFE.

Database	Method	Recognition rates of ten folds and mean (%).										
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Mean
CK+	GSF	92.5	90.2	92.7	92.8	93.2	92.8	90.9	86.1	86.3	82.9	90.1
	HoloNet with SoftMax	95.0	95.1	92.7	97.6	95.5	95.2	93.2	86.1	86.3	87.8	92.5
	The proposed network (λ_C=1e-4, λ_F=1e-4)	**97.5**	**100.0**	**97.6**	**100**	**97.7**	**100.0**	**100.0**	**91.7**	**90.9**	**97.6**	**97.35**
JAFFE	GSF	81.0	81.0	68.2	**76.2**	61.9	**81.8**	**100.0**	**57.1**	**95.2**	**72.7**	77.5
	HoloNet with SoftMax	90.5	95.2	**81.8**	**76.2**	66.7	72.7	90.5	**57.1**	90.5	59.1	77.9
	The proposed network (λ_C=1e-4, λ_F=1e-4)	**95.2**	**100.0**	**81.8**	**76.2**	**95.2**	68.2	**100.0**	**57.1**	**95.2**	68.2	**83.6**

TABLE VI: The effects of the HNwFL and fusion network.

Database	Different network structures and the average recognition rates (%).		
	MI+MII	**MIII**	**MI+MII+MIII**
CK+	96.63	92.05	**97.35**
JAFFE	83.10	78.87	**83.57**
FER2013	61.3	58.85	**61.86**

and the fusion network are beneficial to the recognition rate improvement and HNwFL works much better when all three networks are used.

Considering the runtime cost, the proposed algorithm requires an additional feature extraction process, in addition to the DNN learning. However, the runtime cost of the hand-crafted feature extraction is negligible compared with DNN learning. In the testing stage, the runtime cost of the proposed algorithm includes an additional matrix multiplication with the stored projection matrix for GSF feature extraction in equation (4).

B. Comparison with Other Algorithms

To evaluate the performance of the proposed algorithm with other algorithms, Table VII,VIII list the recognition rates of the proposed and state-of-the-art algorithms on the JAFFE and CK+ databases. One can observe from Table VII that the proposed algorithm ranks the 2nd among the considered algorithms, which justifies its competitive performance. For the CK+ database, it was reported in [39] that peak-piloted DNN achieves a recognition rate of 99.3%. However, both the peak and non-peak (neutral) expressions were needed. Compared with the fine-tuning method [2], the proposed approach did not use any geometry feature or temporal video information. However, Table VIII shows that the proposed algorithm achieves the best recognition rate of 97.35% on the database under the same setting.

TABLE VII: Comparison of different algorithms on JAFFE database.

Algorithm	Subjects	Protocol	Recog. rate (%)
Subclass Discriminant [33]	10	10-fold	49.47
KCCA [34]	10	10-fold	77.05
Weighted LDA [35]	10	10-fold	58.53
Information Projection [36]	10	5-fold	83.18
Classifier Selection [37]	10	10-fold	**85.92**
Ours	10	10-fold	83.57

TABLE VIII: Comparison of different algorithms on CK+ database.

Algorithm	Subjects	Protocol	Recog. rate (%)
Margin Projection [19]	100	5-fold	89.2
Radial Feature [20]	94	10-fold	91.51
AU Network [21]	118	10-fold	92.05
DNN [22]	106	5-fold	93.2
Patch Weighting [38]	106	10-fold	94.09
Fine Tuning [2]	106	10-fold	97.25
Peak-Piloted DNN [39]	106	10-fold	97.3
Ours	106	10-fold	**97.35**

IV. CONCLUSIONS AND FUTURE WORKS

This work proposed a general framework for embedding a hand-crafted feature into a deep network for feature learning. The proposed framework learns a deep feature with the guidance of a hand-crafted feature using deep metric learning. The feature is then integrated with the hand-crafted feature using a fusion network for the final recognition. The experimental results on CK+, JAFFE and FER2013 databases show that the proposed algorithm achieved better performance than the original hand-crafted feature and the feature learned without using the proposed feature loss. Better performance than state-of-the-art approaches has also been observed.

However, there is large room for further improvement. First, more effective hand-crafted features, such as dense SIFT [1] with bag of words will be embedded into more network layers to test the performance of hand-crafted feature guidance. Second, the proposed feature learning will be applied to other applications like face recognition. Third, consider the limitation of the employed network and hand-crafted feature, more effective network and hand-crafted feature for the databases in the wild, such as FER2013 will be investigated in our future work. Lastly, the proposed algorithm will employ larger database to train network parameters with better generalization ability.

V. ACKNOWLEDGMENTS

The work was supported by Natural Science Foundation of China under grands no. 61602315 and 61672357, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20170302153827712, the Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University, the School Startup Fund of Shenzhen University under grands no. 2018063.

REFERENCES

[1] R. T. Ionescu, M. Popescu, and C. Grozea, "Local learning to improve bag of visual words model for facial expression recognition," in *Workshop on Challenges in Representation Learning, International Conference on Machine Learning*, 2013.

[2] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *IEEE International Conference on Computer Vision*, 2016, pp. 2983-2991.

[3] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring bag of words architectures in the facial expression domain," in *European Conference on Computer Vision: Workshops and Demonstrations*, 2012, pp. 250-259.

[4] A. Zadeh, T. Baltrusaitis, and L. P. Morency, "Convolutional experts constrained local model for facial landmark detection," in *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2017, pp. 2051-2059.

[5] H. Pan and H. Jiang, "Learning convolutional neural networks using hybrid orthogonal projection and estimation," in *arXiv:1606.05929v4*, 2016.

[6] H. Qian, Y. Zhang, and C. Liu, "Vehicle classification based on the fusion of deep network features and traditional features," in *Seventh International Conference on Advanced Computational Intelligence*, 2015, pp. 257-262.

[7] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild," in *International Conference on Multimodal Interaction*, 2014, pp. 494-501.

[8] R. Paul, S. H. Hawkins, L. O. Hall, D. B. Goldgof, and R. J. Gillies, "Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2016, pp. 2570-2575.

[9] S. P. Suggu, K. N. Goutham, M. K. Chinnakotla, and M. Shrivastava, "Hand in glove: deep feature fusion network architectures for answer quality prediction in community question answering," in *International Conference on Computational Linguistics (COLING)*, 2016, pp. 1429-1440.

[10] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, "Combining deep learning and hand-crafted features for skin lesion classification," in *International Conference on Image Processing Theory Tools and Applications (IPTA)*, 2016, pp. 1-6.

[11] O. Araque, I. Corcuera-Platas, J. F. Snchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, no. C, pp. 236-246, 2017.

[12] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *IEEE International Conference on Computer Vision Workshop*, 2015, pp. 19-27.

[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision, ECCV*, 2014, pp. 818-833.

[14] L. L. Shen and L. Bai, "Gabor feature based face recognition using kernel methods," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 170-176.

[15] F. Juefei-Xu, V. N. Boddeti, and M. Savvides, "Local binary convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, 2016, pp. 499-515.

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: deep hypersphere embedding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 212-220.

[18] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017) Workshop*, 2017, pp. 522-531.

[19] S. Nikitidis, A. Tefas, and I. Pitas, "Maximum margin projection subspace learning for visual data analysis," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4413-4425, 2014.

[20] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognition*, vol. 45, no. 1, pp. 80-91, 2012.

[21] M. Liu, S. Li, S. Shan, and X. Chen, "AU-aware deep networks for facial expression recognition," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1-6.

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

[23] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition," *Pattern Analysis & Applications*, vol. 9, no. 2-3, pp. 273-292, 2006.

[24] K. Yan, Y. Chen, and D. Zhang, "Gabor surface feature for face recognition," in *Proceedings of First Asian Conference on Pattern Recognition*, 2011, pp. 288-292.
[25] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, 1999.
[26] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: towards robust emotion recognition in the wild," in *ACM International Conference on Multimodal Interaction*, 2016, pp. 472-478.
[27] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, pp. 1-8.
[28] R. Collobert, "Torch Tutorial," Institut Dalle Molle d'Intelligence Artificielle Perceptive Institute, 2002.
[29] T. Kanade, Y. Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46.
[30] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 1998, pp. 200.
[31] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, and B. Hamner, et al., "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*, 2013, pp. 117-124.
[32] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476-3483.
[33] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 28, no. 8, pp. 1274-1286, 2006.
[34] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 233, 2006.
[35] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted Piecewise LDA for Solving the Small Sample Size Problem in Face Verification," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 506-519, 2007.
[36] H. Wang, S. Chen, Z. Hu, and W. Zheng, "Locality-preserved maximum information projection.," *IEEE Transactions on Neural Networks*, vol. 19, no. 4, pp. 571-585, 2008.
[37] M. Kyperountas, A. Tefas, and I. Pitas, "Salient feature and reliable classifier selection for facial expression classification," *Pattern Recognition*, vol. 43, no. 3, pp. 972-986, 2010.
[38] W. Xie, L. Shen, M. Yang, and Z. Lai, "Active AU based patch weighting for facial expression recognition," *Sensors*, vol. 17, no. 2, pp. 275, 2017.
[39] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European Conference on Computer Vision*, 2016, pp. 425-442.