Mir:Multi-task Intra-Representation Learning For Face Classification

Senlin Cheng
Beijing University of Posts and Telecommunications
Beijing, China
972705994@qq.com

Liutong Xu
Beijing University of Posts and Telecommunications
Beijing, China
xliutong@bupt.edu.cn

Pengfei Wang
Beijing University of Posts and Telecommunications
Beijing, China
wangpengfei@bupt.edu.cn

Abstract-Despite significant recent advances in the field of face recognition, implementing face recognition efficiently on a small scale datasets and how to learn a better intra-representation of the face present serious challenges. In this paper we present a new method called MIR, that directly learns similarity feature between face and face and uses characteristic as a new representation to do classification. Our method have a better performance with the characteristic than original features in different datasets(LFW, Youtube Faces(YTF) and MegaFace challenge). Our method **describe** the effects of face classifications after learning similarity and demonstrate that multitasking get a better accuracy than single task learning tasks.

Keywords-multitask learning; intra-representation; face recognition; deep learning

I. INTRODUCTION

Face recognition is developing very fast in the last couple of years. There are many new methods appearing and have near perfect results in the LFW Data-sets and some similar face database, especially in deep-learning. For example, face++ [1], deep-face [2], FR+FCN [3], We call this task Face-Verification, although the accuracy is high in the data-set, is low in reality. All we know is sometimes we need to know how many people in this image and who they are in many applications especially using in the monitoring. Face recognition is widely used in Security monitoring, tracking and so on.

Now there are some methods like Face-Net[6], they use triplet loss to train so that they can minimize the distance between the same class people and maximize the distance from different classes, but there are some disadvantages, firstly, they did not consider the common feature from all people. Secondly, they did not use the character feature of different people to do classifications. Based on these disadvantages, in this paper, we proposed a multitask learning approach for many classification task. We think all people have its similarity and character, if we can extract the commonality from people, we use the character to classification, In theory, we can get a good result. Firstly, we use the Face-net [6] to generate 128-dimension vector to represent a face, The training stage use Triplet Loss to minimize the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity. Next stage, we put the vector into our small-size neural network, we learn the face-similarity and character so that we can do classification.

II. RELATED WORK

A. Traditional face recognition

Early in face recognition, use Bayesian face [10], use template matching method, For example, [7] use human facial features to establish a three-dimensional adjustable model framework. they use all the nodes in the face and measure the distance between eye, chin and cheekbones. But the process is very complicated and the accuracy is low. The classification of HMM and Singular Value Decomposition[12] is a method to do feature extraction. Generally, an image (length H)is regarded as an N*M matrix and its singular value is taken as a feature of face recognition. they use the sampling window to overlap the same picture. The singular values are grouped into a set of vectors and this group of vectors are unique. Lades and some people proposed an architecture named Dynamic Link[13] to identify faces. we can get the vector marker from the Gabor wavelet decomposition of the image position and the connection node distance vector of the graph. they use the elastic graph matching method with an accuracy rate of 97.3%.combined with the Compressive Sensing (CS) theory and proposed a sparse Representation Classification (SRC) face recognition method. The basic idea is that test face images can be expressed by several training images. Due to the design of a bridge between test data and training data, this method always achieve good performance, even in the face of the presence of occlusion and other complex situations.

978-1-5386-6565-7/18/$31.00 ©2018 IEEE

B. Deep learning model

Recently, most deep models have been used for face verification or identification, [19] uses mutil-patch deep CNN and deep metric learning, the feature extraction is using the deep convolution neural network in different regions of the face and then decrease dimension to 128, achieved the accuracy of 0.9977 on the LFW dataset and they claim that their approach is the best way in the world. The work of [14] employ a deep network to use semantic feature mapping of all the same label on the face of the trained label to make the distance small between same classes and the distance large between the different classes.they use deep metric learning and it is one of the most commonly used methods for depth learning in the field of face recognition. Using a better objective function can learn more discriminatory facial features. The deep id [15]learns a 160-dimensional vector that represent the face. Deep 10 increase the data set to adopted the performance. There are two ways to increase the data sets, one approach is to collect good data, for example, the CelebFaces data set, another approach is to make the image multi-scales and multi-channels. [16]The deep-id2 network not only considers the classification, but also considers the inter-class gap. The accuracy is about 99.15% in LFW data-set, but they are different with the method that extract the characters and common features what we do. Deep 102+ changes the network structure and also Deep1D2+ is very robust to occlusion. [8] mainly use the data augmentation, collecting large amounts of data is difficult, for the existing public database face images, they synthesis the new face images from the pose, shape, and expression three aspects, it can train In the LFW and DBA data set as good as the millions of face images. [1]collected 5million face pictures for training deep convolution neural network model from the network. the paper propose although the accuracy is very high in the LFW data set, they have some problems in the paper. There is a deviation in the face data set based on the network, most of the collection photos are smile, make-up, young and beautiful, are quiet different from those in the real scene. Test in the real scene ,the accuracy is only 0.66. Another problem is that face image of the angle, light, age and other differences between the role of each other The accuracy in real application is very low . Therefore, the paper puts forward the direction of further research in the future. one is that we can extract the training data from the video, video in the face of screen is close to the reality of the application scene(change the angle, light, expression and so on). [9]they think under natural conditions, because of the angle, light,occlusions, low resolutions and other reasons, there is a big difference from the face images. the paper presents a new depth learning model to face reconstruction that can learn the invisible side of face image. Thus, the model can greatly reduce the difference between individual face images(the same person, different images).

III. 1MPLEMENTATTON

We use MT-CNN and Face-Net before our own work, and multitask learning to improve the accuracy with the extract feature, and this way is simple but effective.

A. lv/TCNN

MTCNN is a deep-learning method of face alignment and generate face as our training data and it is an input of the Face-net.

MTCNN have three stages cascaded framework, firstly, given an image, we initially re-size it to different scales to build an image pyramid, which is the input of the following three-stage cascaded framework. The first stage is P-Net, Obtained the candidate facial windows and their bounding box regression vectors. Then we employ non-maximum suppression(NMS)to merge highly overlapped candidates. The second stage is R-Net, which rejects a large number false candidates, performs calibration with bounding box regression and conducts NMS. The three stage is similar to the second stage which aim is to identify face regions with more supervision. And the network will output five facial landmarks and positions.

B. FACENET

The task of the face-net in the paper is face verification. Face-net use a deep convolution network trained to directly optimize the embedding itself, To train, they use triplets loss to train the network and achieve state-of-the-art face recognition performance using only 128-bytes per face(they tried different dimensions and result is 128-dimensions vector is the best). The triplet loss describe different versions of face embedding that are compatible to each other and allow for direct comparison between each other.

After getting 128-dimension embedding vector, recognition becomes a KNN classification problem and clustering can be achieved using k-means or clustering.

C. OUR OWN WORK

Suppose that we do a two classifications task(for example, we distinguish the sex is male or female,also the multiply classification is the same), if we use logical regression to do classification, it is denoted as

$$\sigma(wx) = \begin{cases} male & if a(wx) > 0.5 \\ female & \text{if } \sigma(wx) < 0.5 \end{cases} \quad (1)$$

where σ is the logical regression,when the result is greater than 0.5, we think it is male, the opposite is the female. further, we think people have some common features(for example, we describe a person, we use the eyes, nose, mouth...), we think these features may reduce the performance of classification. it is denoted as

$$a(wx+b) > 0.5 \quad (2)$$

If b is large(for example,a=100), the result is always larger than 0.5 so that classification is wrong. we assume the b is the common features.

$$\begin{cases} \vec{x} \bullet \vec{w} = b \\ \sigma(\vec{x} \bullet \vec{w}) = \sigma(\vec{x}_{\text{per}} \bullet \vec{w}_1 + \vec{x}_{com} \bullet \vec{w}_2) \end{cases} \quad (3)$$

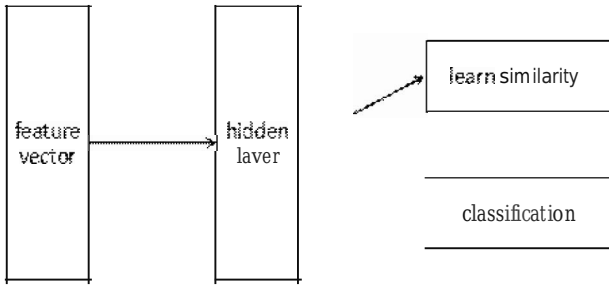

Fig. 1. The network structure of the multitask learning.it is similar with the res-net,we learn the residual and it is the remaining features that we use to do classification task.

where $\vec{x}_{per}$ is the characteristic,$\vec{x}_{com}$ is the similarity features. We assume that we can extract the characteristic from the face embedding featnres.In detail,when we get the face embedding vector from the face-net,we use this to do the multitask learning.When face vector go through the non-linear mapping like sigmoid and it can compress the size of the value between 0-1,we say that if this value is greater than a certain threshold.the data belong to the same class.according to the logical regression.we can divide all faces into a class.that is.we use 0.5 as our threshold here and then we can train the network to learn the similar feature that we need.If we can extract this part of common features.then use the characteristics to do face classification.the results will be more accurate,The multitask is that we achieve two experiments, one is that we think we need to extract the common features ,another is a classification task.We design the network structure.as shown in Fig. 1.

The first layer is the hidden layer, the vector of faces(X) is as the input of the first layer, the dimensions of X are 1*128,we have$x_1 = g(Xrw_1 + b_1)$ as the output of the hidden layer,the g is the activation function,we use the relu activation here.unis the vector of 128*128 dimensions, *bl* is the vector of 1*128 dimensions ,we can get 1*128 dimensions vectors .then the second layer is used to classify the face images to a class so that we can get the common feature .we have $y = a(x_1w_2 + b_2)$, a is the function of sigmoid ,the b_2 is the vector of 128*1 dimensions, b_2 is the vector of 1*1 dimension and we can get 1*1 vector .we think that all the people that there is only one class ,we use

$$y = \sigma(x_1 w_2 + b_2) > \frac{1}{2} \tag{4}$$

We think that y is the common feature that we can extract from all images of the faces. then we use x subtract y to get the character features so that we can do multi-classifications. The last layer is the softmax-layer to do classifications.

We have two ways in the process of the training, One way is that we are fixing a task to learn another task. Another way is that use a joint approach to train, we train two tasks at the same time.

TABLE [
experiments on three data-sets,We have different models with different Datasets ,in which the b in bracket represents the original feature,the a in bracket represents the performance of our way when extract the character feature and do classifications.

DataSet and model	LFW	YTF	MegaFace
DeepID2+[7](b)	98.70%	93.2%	64.21%
DeepID2+[7](a)	99.15%	93.86%	66.23%
FaceNet[6](b)	% 99.65	95.1%	54.85%
FaceNet[6](a)	99.68%	96.72%	55.75%
DeepFace[2](b)	97.35%	91.4%	65.21%
DeepFace[2](a)	97.85%	92.34%	66.78%
BaiDu[20](b)	99.13%	95.6%	67.21%
BaiDu[20](a)	99.18%	96.13%	68.35%

IV. EXPERTMENTS AND RESULTS

In this section, we evaluate the experiments on face recognition. The experiments demonstrate our proposed method is robust on different datasets.can not only improve the accuracy on visual classification, but also boost the performance on visual recognition. We used the datasets/Labeled Face in the Wild(LFW), Youtnbe Faces(YTF) and MegaFace Challenges). We do experiments using Tensorflow,

We evaluate our method on the face recognition task. I.e. given a frame in the video, the evaluation is the precision that predict people who they are and judge if it is right, we can define the evaluation as

$$Ac = \frac{TP}{AP} \tag{5}$$

AP is the number of all people in the picture, TP is the number of all people that is correct when doing classify.

A. The result of different models

This experiment mainly reflects the performance of different models.As we can see in Table I.

B. What we learn in the network

Our network can learn the common features of face images from the network.As can be seen in Fig.2. and Fig.3., We chose 10 photos of two of them in the sample (5 photos of a person) and use PCA to do reducing dimensions, then draw it. After that we use the residual value to do classification, this residual is similar as the residual network[24], the difference is that our network use the subtraction instead of the addition when learning the residual value. the above experiment has shown that our ideas are effective. We find that the distance between different classes become larger after the feature extraction, and which indicates that our network effectively learns our commonality. When training this network, we use the learning rate 0.001 and we add the L2 regularization. We tried different coefficients and choose the best coefficient 0.001. We also do another experiment and make further processing of the data,we centralize the data and do classification again, we can see the performance in the table, after the centralize the data, there

Fig 2. we show the relation when mapping the original feature of two people into two dimensions.

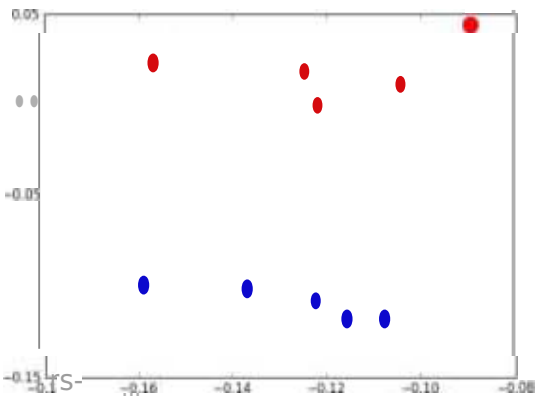

Fig 3, we show the relation when extracting the similar feature and mapping feature of two people into two dimensions.

is a little improvement of the performance.one person and all the photos are collected in the natural state and we think that there will be a better generalization ability.

C. Multi-task vs Single task

We also do a comparative classification experiment with multitask and single task, the result is that mu)titask is better than single task. The single task means that we directly Irai n our Data-sets and then use ordinary FaceNet as a classification and optimize the objective function. We have the result,As shown in Table II.

TABLE II
shows the precision of the two kinds of tasks, the single task is that we train the network with datasets and then we do classification task directly, the result shows th.t the multitask have • better performance than single task.

	Multi-Task	Single-Task
LFW	99.68%	98.71%
YTF	96.72%	93.77%
Mega-Face	55.75%	50.12%

REFERENCES

[1] Erjin Zhou, Zhimin Cao,Qi Yin, Naive-Deep Face Recognition Touching the Limit of LFW Benchmark or Not? 2015 CVPR.

[2] Yaniv Taigman,Ming Yang, MarcAurelio Ranzato,Lior Wolf, DeepFace Closing the G.p to Human-Level Performance in Face Verification 2014 cvra

[3] Zhenyao Zhu,Ping Luo,Xiaogang Wang,Xiaoou Tang,Recover Canonical-View Faces in the Tild with Deep Neural Networks.

[4] A.Samal P.A.Iyengar.Automatic recognition and analysis of human faces and facial expressions:a survey [J), Pattern Recognition199225(1)65-67

[5] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li ,Yu Qiao Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks.ECCV,2016.

[6] F.Schroff, D, Kalenichenko, and J, Philbin. Facenet: A unified embedding for face recognition and clustering, In Proc. CVPR, 2015.

[7] Haddadnia J.Ahmadi MFaez K A Hybrid Learning RBF Neural Network for Human Face Recognition with Pseudo Zernike Moment Invariant *(A)*, Proceedings of the 2002 International Joint Conference on Neural Networks [C].2002111-16.

[S] I Masi, AT 'rnn, T Hassner, Do We Really Need to coueet Millions of Faces for Effective Face Recognition? ECCV,2016.

[9] AT Tran, T Hassner, I Masi, G Medioni Regression Robust and Discriminative 3D Morphable Models with . very Deep Neural Network.

[10] B.Moghaddam,T.Jebra,and A.Pentland. Bayesian face recognation. PR,33:1771-1782,2000.

[11] J, Sivic, M, Everingham, and A. Zisserman. Who are you? learning person specific classifiers from video.In Proc.CVPR, 2009.

[12] Wenyi Zhao. Robust image based 3D face recognition [D), PhD, Thesis. University of Maryland,1999.

[13] Martin Lades,Student Member, IEEE, Jan C, Vorbruggen, Member, IEEE, Joachim Buhmann,Member, IEEE, Jorg Lange, Christoph v.d. Malsburg,Rolf P.Wurtz, and Wolfgang Konen Distortion Invariant Object Recognition in the Dynamic Link Architecture IEEE Transactions on Computers, Vol.42, No,3, MARCH 1993.

[14] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, Silvio Savarese; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp, 4004_4012

[15] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes[C]//Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE,2014:1891-1898.

[16] Y Sun , D Liang ,X Wang , X Tang DeepID3:Face Recognition with Very Deep Neural Networks.

[17] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective and robust[J]. arXiv preprint arXiv:1412.1265, 2014.

[18] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi M., Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence, 3\ (2):210227, 2009.

[19] Jingtuo Liu Yafeng Deng Chang Huang Targeting Ultimate Accuracy Face Recognition via Deep Embedding

[20] J, Liu, Y, Deng, and C, Huang, Targeting ultimate accuracy:Face recognition via deep embedding arXiv preprint:1506.07310, 2015. 3, 7