Cross-Modality Multi-Task Deep Metric Learning for Sketch Face Recognition

Yujian Feng
College of Automation
Nanjing University of Posts and Telecommunications
Nanjing, China
18676148636@163.com

Fei Wu*
College of Automation
Nanjing University of Posts and Telecommunications
Nanjing, China
wufei_8888@126.com

Qinghua Huang
School of Science
Nanjing University of Posts and Telecommunications
Nanjing, China
hqh@njupt.edu.cn

Xiao-Yuan Jing
College of Computer
Wuhan University
Wuhan, China
jingxy_2000@126.com

Yimu Ji
School of Computer Science
Nanjing University of Posts and Telecommunications
Nanjing, China
jiym@njupt.edu.cn

Jian Yu
College of Automation
Nanjing University of Posts and Telecommunications
Nanjing, China
18351927750@163.com

Feng Chen
College of Automation
Nanjing University of Posts and Telecommunications
Nanjing, China
chenfeng1271@gmail.com

Lu Han
College of Automation
Nanjing University of Posts and Telecommunications
Nanjing, China
hanl@njupt.edu.cn

Abstract—**Sketch face recognition is to match face sketch images to photo images. The main challenge of it lies in cross-modality differences. To address this challenge, a variety of methods were proposed to bridge cross-modality gap of different modalities. Specially, common subspace-based methods have achieved great performance in this task. These methods enable the data of different modalities to be comparable by mapping this data into a new and common subspace. However, the problem of non-linear distribution of samples from different modalities has not been well solved by these methods. In this paper, we propose a cross-modality multi-task deep metric learning (CMTDML) approach to address this problem. Firstly, we design a two-channel neural network to extract non-linear features of photo modality and sketch modality, and the parameter sharing characteristics can reduce the differences of features between different modalities. Secondly, we develop the loss function to constrain the features in common space, where intra-class compactness and inter-class separability of features are promoted. In extensive experiments and comparisons with the state-of-the-art methods, the CMTDML approach achieves marked improvements in most cases.**

Index Terms—**Sketch face recognition, Cross-modality gap, Multi-task learning, Deep metric learning, Common subspace learning.**

I. Introduction

In the face recognition community, it is a challenging problem to distinguish faces in two modalities of sketch images and photo images. In real-world applications, sketch face recognition has been widely applied in law enforcement agencies. For example, there is a scene in which the picture of the suspect is very blurred and requires the artist to draw a sketch. When the police get these sketches, they can quickly narrow down the range of suspects. However, due to the large modality gap between mugshot photos and face sketches, sketch-based face recognition remains a challenging topic in the community [1], [2].

Because of the great discrepancies between heterogeneous face images, conventional homogeneous face recognition methods perform poorly by directly identifying the probe image (face sketch or photo) from gallery image (face photo or sketch). Some heterogeneous face recognition methods have been proposed. Existing methods can be generally divided into three categories: synthesis-based methods [1]–[5], feature descriptor-based methods [6]–[8], and common subspace-based methods [9]–[13]. Synthesis-based methods transform the data of one modality into another by synthesizing. Once

* Corresponding author

978-1-7281-4094-0/19/$31.00 ©2019 IEEE

the synthesized photos are generated from other images, conventional face recognition algorithms can be applied directly. Feature descriptor-based methods represent face images with local feature descriptors, and these descriptors can be utilized for face recognition. In this paper, we design a deep common subspace-based method to excavate the discriminant non-linear features from different modalities.

Common subspace-based methods aim to transform different modalities into a common subspace to reduce differences of sketch images and photo images. [9] proposes a discriminant feature extraction method to translate heterogeneous features into the same feature space. [10] utilizes canonical correlation analysis for cross-modality matching. [11] applies the partial least squares (PLS) method to linearly map images in different modalities to a common linear subspace. [12] proposes a multi-view discriminant analysis (MvDA) method to obtain a common space for multiple views by optimizing both inter-view and intra-view Rayleigh quotient. [13] proposes a cross-modality metric learning (CMML) method to learn a discriminative latent space. However, these methods do not take into account the non-linear distribution of samples. Therefore, these methods may not conduct a salient and discriminative feature extractor.

In recent years, numerous metric learning (ML) methods have been proposed in computer vision field. The aim of metric learning is to learn a distance function to measure the similarity between samples. However, most traditional metric learning methods usually learn linear mapping to project samples into a new feature space, which is affected by the non-linear relationships of different modalities. Therefore, deep metric learning (DML) methods have been proposed to learn nonlinear features [14]–[16]. [14] proposes a discriminative deep metric learning method for face verification. [15] proposes a deep nonlinear metric learning method by using a deep independent subspace analysis network. [16] proposes a DML method with a Siamese deep neural network to learn a similarity metric from image pixels directly for person reidentification. Inspired by these deep metric learning methods, in this paper, we propose a cross-modality multi-task deep metric learning (CMTDML) approach for sketch face recognition. The main contributions of this paper are summarized as follows:

1) We design a two-channel neural network to extract non-linear features from two modalities, and reduce the difference between modalities by sharing the parameter of the network.
2) We utilize a multi-task learning mechanism to perform intra-modality discriminant analysis and inter-modality discriminant analysis. We specifically design an intra-modality discriminant analysis loss function, which can make full use of information of sample labels. This loss function also makes the feature distribution uniform within modality. It helps to reduce the gap between modalities. In addition, we design an inter-modality discriminant analysis loss function, which makes intra-class features more compact and inter-class features more separated.
3) Our method is evaluated on two benchmarks, CUFS [2] and CUFSF [2], [22]. And state-of-the-art results are achieved by CMTDML in these two databases.

The rest of the paper is organized as follows. Details of the proposed methods are given in Section II. In Section III, experiments and results are presented together with discussions. the Study is concluded in Section IV. Finally, acknowledgments in Section V.

II. Our Method

A. Notations

Suppose $X = \{x_i|i = 1, 2, \cdots, k_x\}$ is a set of training samples, k_x is the number of the training samples. x_i and x_j are the i-th sample and the j-th sample in X, respectively. We pair each two samples in the sample set to attain different combinations of all samples. Each pair of samples contains sketch modality and photo modality, which are defined as s and p, respectively. x_i^p is the i-th sample of photo modality, and x_j^s is the j-th sample of sketch modality. $P_{x_i^p}$ is the probability of x_i^p , $P_{x_j^s}$ is the probability of x_j^s . If prediction of category is the same as ground truth, $y = 1$, otherwise $y = 0$.

B. Framework

The framework of CMTDML is shown in Fig. 1. There are two tasks including intra-modallity discriminant analysis and inter-modality discriminant analysis. We design a two-channel neural network to extract features of different modalities, and reduce the difference between modalities by taking advantage of parameter sharing mechanism. For the first task, we fully utilize the label information of samples to excavate the relationship between the non-linear features of samples. For the second task, we use a distance function to constraint features in common space to promote intra-class compactness and inter-class separability.

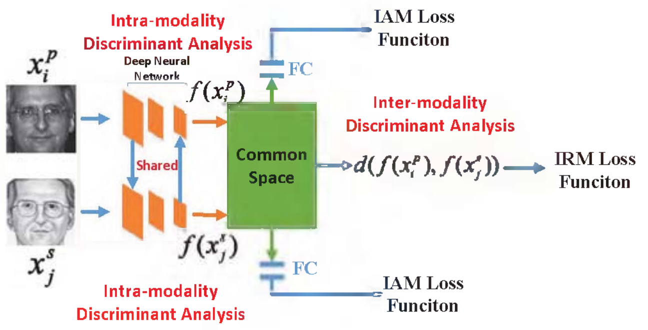

Fig. 1: The framework of CMTDML. It contains two tasks including intra-modality discriminant analysis and inter-modality discriminant analysis. Deep neural network without the last fully connect layer(FC) is used as the feature extractor.

C. Loss Function of CMTDML

Our approach includes the intra-modality discriminant analysis (IAM) loss function and the inter-modality discriminant analysis (IRM) loss function.

1) The IAM Loss Function: For intra-modality discriminant analysis, the IAM loss function is designed to perform the task. For the photo modality, we define L_p loss function showed as follows:

$$\begin{aligned} L_p = & -\sum_{i=1}^{k_x}[y\log(P_{x_i^p}) + (1-y)\log(1-P_{x_i^p})] \\ & + \frac{1}{k_x}\sum_{i=1}^{k_x}(P_{x_i^p} - E(P_{x_i^p}))^2 \end{aligned} \quad (1)$$

$E(\cdot)$ is the mean function. The first term is an cross entropy error based function to make use of label information of samples to extract non-linear features from different modalities. For the second term, we use a distribution constraint on the features to make the distribution of features within the modality more uniform. The distribution constraint is defined on the mean variance of the probability of each category.

At the same time, we define L_s loss function as L_p:

$$\begin{aligned} L_s = & -\sum_{j=1}^{k_x}[y\log(P_{x_j^s}) + (1-y)\log(1-P_{x_j^s})] \\ & + \frac{1}{k_x}\sum_{j=1}^{k_x}(P_{x_j^s} - E(P_{x_j^s}))^2 \end{aligned} \quad (2)$$

We minimize the loss functions L_p and L_s to achieve discriminant analysis within modality. By using second term in Eqs. (1) and (2), the features in common spaces are uniformly distributed for both modalities. It helps to reduce the difference between modalities.

2) The IRM Loss Function: For inter-modality discriminant analysis, we consider intra-class correlation and inter-class difference of different modalities to improve recognition accuracy. For the input x_i^p and x_j^s, $f(x_i^p)$ and $f(x_j^s)$ are the corresponding output through the neural network. The distance of samples x_i^p and x_j^s in common space can be measured by the Euclidean distance between $f(x_i^p)$ and $f(x_j^s)$ as follows:

$$d(f(x_i^p), f(x_j^s)) = ||f(x_i^p) - f(x_j^s)||_2 \quad (3)$$

When the distance in common space is obtained, IRM loss function aims to learn an appropriate deep metric to perform discriminant analysis between modalities. It facilitates intra-class correlation and inter-class separation by minimizing intra-class distance and maximizing the inter-class distance. Specifically, we define IRM loss function showed as follows:

$$\begin{aligned} L_{IRM} = & \sum_{(i,j)\in S} h(d(x_i^p, x_j^s) - \tau) \\ & + \alpha \sum_{(i,j)\in D} h(\tau - d(x_i^p, x_j^s)) \end{aligned} \quad (4)$$

where $h(t) = max(0,t)$ is the hinge loss function. $S = \{(i,j)\}$ and $D = \{(i,j)\}$ represent the indexes of similar pairs and dissimilar pairs, respectively. Obviously, the number of similar pairs is small than dissimilar pairs. By adjusting the value of α, we can will deal with this imbalance problem. τ is a threshold. By minimizing the IRM loss function, the distance $d(x_i^p, x_j^s)$ of intra-class features is expected to be smaller than threshold τ, and the distance of inter-class features is expected to be larger than threshold τ.

III. Experiments

A. Databases and Evaluation Protocols

CUHK (The Chinese University of Hong Kong) Face Sketch database (CUFS) is for research on sketch face synthesis and sketch face recognition, containing 606 faces totally. CUHK Face Sketch FERET Database (CUFSF) contains 1194 persons from the FERET database. For each face, it has a sketch drawn by the artist based on light conditions and shape exaggeration. In order to evaluate our approch on CUFS, we use the usage agreement in [2]. To evaluate our approch on CUFSF, we use the usage agreement in [22], we randomly select 500 subjects as the training set, which is also utilized to generate image pairs. The remaining 694 subjects are used for testing . Fig. 2 gives some examples of two databases.

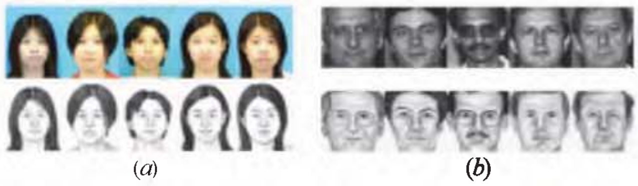

Fig. 2: Examples of face sketch-photos on (a) CUFS database and (b) CUFSF database. The first row shows the original photos and the second row shows the sketches.

B. Experimental Settings

Images are cropped and resized to resolution of the 128×128 and converted to grayscale. We choose three deep neural networks as feature extractor. Our hardware configuration comprises 64-bit computer with Inter i7-8700 CPU, NVIDIA GeForce GTX 1080 Ti. And we use Tensorflow and CUDA 9.0. We train the network on a single NVIDIA GeForce GTX 1080 Ti. τ = 0.2 and α = 0.3 is the best setting on CUFS, τ = 0.5 and α = 0.2 is the best setting on CUFSF. Experimental results in this paper are mean results of 20 random running.

1) Parameter Analysis of τ and α: We take CUFSF database as example and analyze the parameters τ and α. Fig. 3 shows the recognition accuracy of our approach with different value of τ and α from 0.1 to 1.0 with step size 0.1. When the threshold value τ is between 0.4 and 0.6, the experimental result is better and stable. For α, we can achieve better and stable recognition accuracy, when it is set between 0.2 and 0.4. This means that when it is set in this range, the imbalance problem can be alleviated. We can observe a similar phenomenon in the CUFS database.

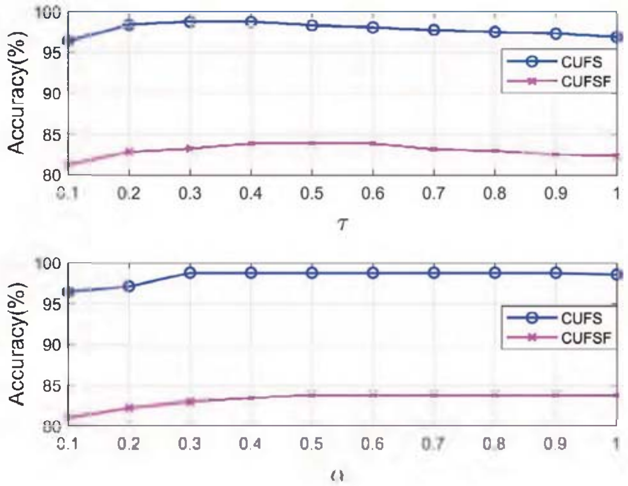

Fig. 3: The influence of τ and α on CUFS database and CUFSF database.

2) *Feature extractor:* The results of our approch utilizing different deep neural networks as feature extractor are shown in Fig. 4. We employ three deep neural networks containing VGG [19], ResNet [20] and DenseNet [21], and compare the accuracy of different rank of each networks. In Fig. 4, the experimental results are better on the DenseNet network. The DenseNet network reduces the phenomenon of vanishing-gradient and enhances the transfer of features that makes better use of features. Therefore, the results prove that the network of DenseNet can extract robust features of different modalities by utilizing a nonlinear function. Similar phenomena can be observed in the CUFS database.

C. Experimental Results

For CUFS database, we evaluate the rank-1 recognition accuracies, which are shown in Table I. CMTDML is compared with some state-of-the-art methods, including PLS [11], CDFE [9], MWF [4] and Fast-RSLCR [5]. From the table, Fast-RSLCR achieves good result, and our approach performs the best and improves accuracy by 0.43% (= 98.78% - 98.35%).

TABLE I: Average rank-1 recognition accuracy (%) of the state-of-the-art methods and our approach on the CUFS database.

Methods	CDFE [9]	MWF [4]	PLS [11]	Fast-RSLCR [5]	CMTDML
Rank-1(%)	75.00	92.13	93.60	98.35	98.78

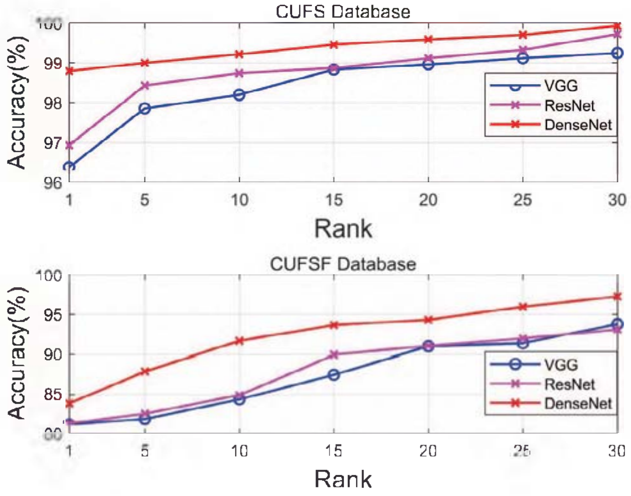

Fig. 4: The results of different feature extractors on CUFS database and CUFSF database.

TABLE II: Average rank-1 recognition accuracy (%) of the state-of-the-art methods and our approach on the CUFSF database.

	Methods	Rank-1 recognition(%)
Synthesis-based Methods	MRF [2]	46.03
	MWF [4]	74.00
	Fast-RSLCR [5]	75.94
Feature Descriptor-based Methods	C-DFD [7]	74.60
	CDFL [6]	81.30
Common Subspace-based Methods	CDFE [9]	47.60
	PLS [11]	51.00
	MvDA [12]	55.00
	CMML [13]	80.00
Deep Learning Methods	VGG [17]	45.82
	SeetaFace [18]	16.57
Our approach	CMTDML	83.86

On the CUFSF database, we compare our CMTDML approach with state-of-the-art methods, and the results are shown in Table II. For synthesis-based methods of MRF [2], MWF [4] and Fast-RSLCR [5], the distortion problem of synthesis because of the CUFSF database is affected by light conditions and exaggerated shaped, results to degrade the performance of sketch face recognition. For feature descriptor-based methods including CDFL [6] and C-DFD [7], and common subspace-based methods including PLS [11], MvDA [12], CDFE [9] and CMML [13], there exists some room to improve their recognition accuracy. For the deep learning methods of VGG [17] and SeetaFace [18], these models are trained on visible photos rather than sketches, and the performance is relatively poor, which indicates that directly using existing deep learning methods can not bring good sketch face recognition accuracy. From the table, our approach can achieve the best recognition performance among compared methods. And it improves the rank-1 accuracy by 2.56% (= 83.86% - 81.30%). The reasons for the improvement are: our approach makes effort to explore the non-linear relationship of samples, effectively reduces the gap between modalities, and skilfully performs discriminant analysis from intra-modality and inter-modality aspects.

IV. Conclusions

In this paper, a novel face sketch recognition approach is proposed to explore non-linear relationship between samples. The cross-modality gap is effectively reduced through the parameter sharing mechanism. The discrimination information is fully explored by performing intra-modality discriminant analysis and the inter-modality discriminant analysis. Experimental results on two databases demonstrate the effectiveness and superiority of the proposed method.

V. Acknowledgments

The work in this paper was supported by the National Natural Science Foundation of China (Nos. 61702280, 61872196), National Postdoctoral Program for Innovative Talents (No. BX20180146), Natural Science Foundation of Jiangsu Province (No. BK20170900), National Key R&D Program of China (Nos. 2017YFB1401302, 2017YFB0202200), NSFC-Key Project of General Technology Fundamental Research United Fund (No. U1736211), and Scientific Research Starting Foundation for Introduced Talents in NJUPT (NUPTSF, No. NY217009).

References

[1] X. Tang. and X. Wang: Face sketch recognition. IEEE Transactions on Circuits and Systems for Video Technology 14(1), 50-57, 2004.

[2] X. Wang. and X. Tang: Face photo-sketch synthesis and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11), 1955-1967, 2009.

[3] S. Zhang, *et al.*: Robust face sketch style synthesis. IEEE Transactions on Image Processing 25(1), 220-232, 2016.

[4] H. Zhou, Z. Kuang, K. Y. K. Wong: Markov weight fields for face sketch synthesis. IEEE Conference on Computer Vision and Pattern Recognition 2012, pp. 1091-1097.

[5] N. Wan, X. Gao, J. Li: Random sampling for fast face sketch synthesis. Pattern Recognition 76, 215 - 227, 2018.

[6] Y. Jin, J. Lu, Q. Ruan: Coupled discriminative feature learning for heterogeneous face recognition. IEEE Transactions on Information Forensics and Security 10(3), 640-652, 2015.

[7] Z. Lei, *et al.*: Learning discriminant face descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(2), 289-302, 2013.

[8] W. Zhang, X. Wang, X. Tang: Coupled information-theoretic encoding for face photo-sketch recognition. IEEE Conference on Computer Vision and Pattern Recognition 2011, pp. 513-520.

[9] D. Lin. and X. Tang: Inter-modality face recognition. European Conference on Computer Vision 2006, pp. 13-26.

[10] D. Yi, *et al.*: Face matching between near infrared and visible light images. International Conference on Biometrics 2007, pp. 523-530.

[11] A. Sharma. and D. W. Jacobs: Bypassing synthesis: PLS for face recognition with pose low-resolution and sketch. IEEE Conference on Computer Vision and Pattern Recognition 2011, pp. 593-600.

[12] M. Kan, *et al.*: Multi-view discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(1), 188-194, 2016.

[13] A. Mignon. and F. Jurie: CMML: A new metric learning approach for cross modal matching. Asian Conference on Computer Vision 2012, pp. 1-14.

[14] J. Hu, J. Lu, Y. P. Tan: Discriminative deep metric learning for face verification in the wild. IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 1875-1882.

[15] X. Cai, *et al.*: Deep nonlinear metric learning with independent subspace analysis for face verification. ACM International Conference on Multimedia 2012, pp. 749-752.

[16] D. Yi, *et al.*: Deep metric learning for person re-identification. International Conference on Pattern Recognition 2014, pp. 34-39.

[17] O. M. Parkhi, A. Vedaldi, A. Zisserman: Deep face recognition. The British Machine Vision Conference 2015, pp. 1-12.

[18] X. Liu, *et al.*: VIPLFaceNet: An open source deep face recognition SDK. Frontiers of Computer Science 11(2), 208-110, 2017.

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

[20] K. He, *et al.*: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 770-778.

[21] G. Huang,*et al.*: Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 4700-4708.

[22] W. Zhang, X. Wang and X. Tang. Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition. IEEE Conference on Computer Vision and Pattern Recognition 2011, pp. 513-520