Deep Metric Learning with Triplet-Margin-Center Loss for Sketch Face Recognition

SUMMARY Sketch face recognition is to match sketch face images to photo face images. The main challenge of sketch face recognition is learning discriminative feature representations to ensure intra-class compactness and inter-class separability. However, traditional sketch face recognition methods encouraged samples with the same identity to get closer, and samples with different identities to be further, and these methods did not consider the intra-class compactness of samples. In this paper, we propose triplet-margin-center loss to cope with the above problem by combining the triplet loss and center loss. The triplet-margin-center loss can enlarge the distance of inter-class samples and reduce intra-class sample variations simultaneously, and improve intra-class compactness. Moreover, the triplet-margin-center loss applies a hard triplet sample selection strategy. It aims to effectively select hard samples to avoid unstable training phase and slow converges. With our approach, the samples from photos and from sketches taken from the same identity are closer, and samples from photos and sketches come from different identities are further in the projected space. In extensive experiments and comparisons with the state-of-the-art methods, our approach achieves marked improvements in most cases.

key words: sketch face recognition, triplet loss, center loss, hard triplet sample selection

1. Introduction

In a real-world application, 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 methods perform poorly by directly identifying the probe images (face sketch or photo) from the gallery images (face photo or sketch). Thus, to accurately and quickly retrieve law enforcement face data sets or surveillance camera lenses, it is necessary to propose an automatic sketch face recognition method.

Traditional sketch face recognition methods mainly include three categories: feature descriptor-based methods [1], [2], synthesis-based methods [3], [4] and common subspace-based methods [5], [6]. Feature descriptor-based methods represent face images with local feature descriptors, these encoded descriptors can then be utilized for recognition. Synthesis-based methods map the face image of one modality to another one through photo-sketch synthesis. Common subspace-based methods aim to transform different modalities into a common subspace to reduce differences of sketch images and photo images.

Recently, deep learning technology has made great progress in the field of sketch face recognition by learning potential embedded information from face images data. Galea et al. [7] varied facial attributes and automatically synthesized a new large set of images. Mittal et al. [8] presented a novel algorithm for matching composite sketches with photographs using transfer learning with deep learning representation. Deng et al. [9] proposed a residual compensation network to reduce the modal discrepancy.

In order to learn more robust and discriminative features, deep metric learning has been widely adopted in multiple fields. Cheng et al. [10] proposed a novel convolutional neural network model under the triplet framework to pull the instances of the same identity closer, and push the instances belong to different identities farther for each other. He et al. [11] proposed a loss to learn multiple class centers and required that the distances between samples and centers from the same class are closer than those from different classes. Wan et al. [12] designed a three-channel convolutional neural network architecture in which triplet loss is adopted in order to learn discriminative features and intra-class variations.

Inspired by these deep metric learning methods, in this paper, we propose a triplet-margin-center loss for sketch face recognition based on three-channel convolutional neural network. In three-channel convolutional neural network, a triplet example consists of a sample as anchor sample (a sketch image), a positive sample (same identity as the anchor sample) and a negative sample (different identity from anchor sketch). Then, we design a triplet-margin-center loss to ensure intra-class compactness and inter-class separability. This loss can automatically select hard samples, and is instrumental in training process because no hard sample needs to be selected manually. The main contributions of this paper are summarized as follows:

1. A triplet-margin-center loss function is not only to enlarge inter-class differences and reduce intra-class variations, but also to learn the center of each class and pull features of the same class to the corresponding center
more closer.

2. We propose a hard triplet sample selection strategy to select informative hard samples. We introduce a threshold, when the distance between the negative sample and the center (the center of an anchor sample and the positive sample) is smaller than the threshold, then the negative samples as hard samples are selected.

3. Our approach is evaluated on two benchmarks, CUFSEF [13, 14] and IIIT-D [15]. Experimental results indicate that the proposed sketch face recognition method outperforms the existing approaches.

2. Our Method

2.1 Notations

Suppose \( X = \{(x_i)\}_{i=1}^n \) is a set of training samples. \( n \) is the number of training samples. \( x_i \) and \( x_j \) are the \( i \)-th sample and the \( j \)-th sample in \( X \), respectively. We pair each three samples in the sample set to attain different combinations of all samples, and obtain triplet sample pairs. Each triplet sample pair contains \( x_i^s \), \( x_i^p \), and \( x_j^p \), where sketch image and photo image are defined as \( s \) and \( p \) respectively. \( x_i^s \) is the \( i \)-th sample of sketch image as an anchor sample, \( x_i^p \) is the \( i \)-th sample of photo image as a positive sample, \( x_j^p \) is the \( j \)-th sample of photo image as a negative sample. \( c_i \) is the center of class corresponding to \( x_i^s \). The sketch face database has only a pair of samples (sketch image and photo image) for the same identity person, and \( c_i \) can be obtained by the mean of \( x_i^s \) and \( x_i^p \).

2.2 Framework

We design a three-channel neural network in Fig. 1. Firstly, we obtain triplet sample pairs by the pairing of sketches and photos. Each triplet sample pair consists an anchor sketch \( x_i^s \), a positive photo \( x_j^p \) (same identity as the anchor sketch \( x_i^s \)) and a negative photo \( x_{ij}^p \) (different identity from anchor sketch \( x_i^s \)). Secondly, when the triplet sample pairs are input into the three-channel network, we learn the corresponding features \( f(x_i^s) \), \( f(x_i^p) \) and \( f(x_{ij}^p) \) respectively. \( f(c_i) \) is the center of \( f(x_i^s) \) and \( f(x_i^p) \). \( f(x_{ij}^{p,s}) \) is the general representation of \( f(x_i^p) \) or \( f(x_i^s) \), \( d(f(x_{ij}^{p,s}), f(c_i)) \) denotes \( d(f(x_i^p), f(c_i)) \) or \( d(f(x_i^s), f(c_i)) \). Then, we require the distance \( d(f(x_{ij}^{p,s}), f(c_i)) \) between \( f(x_{ij}^{p,s}) \) and \( f(c_i) \) to be smaller than the distance \( d(f(x_i^p), f(c_i)) \). Moreover, we design triplet-margin-center loss for the sake of learning more discriminative features, and the hard negative examples are effectively selected for improving the stability of training.

2.3 Triplet-Margin-Center Loss Function

Intuitively, triplet-margin-center loss encourages the distances between samples from the same class to be smaller than those from different classes by a threshold \( m \). Specifically, we define triplet-margin-center loss function as follows:

\[
L = L_t + \lambda L_{mc}
\] (1)

where \( \lambda \) is a hyper-parameter to balance the loss function \( L_t \) with the loss function \( L_{mc} \).

The loss function \( L_t \) is designed to minimize the intra-class distances of the learned deep features as well as to maximize the inter-class distances of deep features simultaneously. We define the loss function \( L_t \) as follows:

\[
L_t = \sum_{i=1}^n \max(0, m + d(f(x_{ij}^{p,s}), f(c_i)) - d(f(x_i^p), f(c_i)))
\] (2)

where \( x_{ij}^{p,s} \) is a positive sample \( x_i^p \) or an anchor sample \( x_i^s \). We utilize a threshold \( m \) between \( f(x_{ij}^{p,s}) \) and \( f(c_i) \) versus \( f(x_i^p) \) and \( f(c_i) \). By minimizing the \( L_t \) loss function, \( d(f(x_{ij}^{p,s}), f(c_i)) \) is pushed to be smaller and \( d(f(x_i^p), f(c_i)) \) is pushed to be larger than \( d(f(x_{ij}^{p,s}), f(c_i)) + m \). Then, the distance of \( d(f(x_i^p), f(c_i)) \) will be smaller, and the sample \( x_i^p \) will be close to the center \( c_i \). The distance of \( d(f(x_{ij}^{p,s}), f(c_i)) \) will also be smaller, and the sample \( x_{ij}^{p,s} \) will be close to the center \( c_i \). Therefore, intra-class samples \( x_i^p \) and \( x_i^s \) will be compact with center \( c_i \).

However, the number of triplet sample pairs grows cubically when the training dataset gets larger, which usually results in training unstable and converges slowly. It is crucial to select the hard samples that contribute to training effectively. Thus, we define the loss function \( L_{mc} \) follows:

\[
L_{mc} = \sum_{i=1}^n \sum_{j=1, i \neq j}^n (\|x_{ij}^p - c_i\| - \alpha_{ij})
\] (3)

where \( \alpha_{ij} \) is a margin of our loss function. The loss function \( L_{mc} \) includes hard samples \( x_i^p \) that satisfies \( \|x_{ij}^p - c_i\| < \alpha_{ij} \). Under this condition, the hard sample of negative photo \( x_{ij}^p \) can be effectively selected. For our hard sample selection strategy, \( \alpha_{ij} \) is set as 1.0 in our experiments.

In the triplet-margin-center loss function, class centers \( c_i \) should be updated as the deep features captured. It is impractical and ineffective to update the class centers among the whole training set. Thus, we update the center in mini-batch. In each iteration, the centers would be updated by the samples in the mini-batch.

All in all, we design the triplet-margin-center loss function, which not only makes intra-class features be more
compact and inter-class features be more separable, but also can effectively select hard samples to avoid the slow convergence in training.

3. Experiments

3.1 Databases and Evaluation Protocols

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. To evaluate our approach on CUFSF, according to the usage agreement [13], we randomly select 500 subjects as the training set. The remaining 694 subjects are used for testing.

IIIT-Delhi Viewed Sketch database includes a total of 238 sketch-digital image pairs. The sketches are drawn by a professional sketch artist for digital images collected from different sources. It comprises of 67 sketch-digital image pairs from the FG-NET aging database, 99 sketch-digital image pairs from Labeled Faces in Wild database, and 72 sketch-digital image pairs from the IIIT-D student and staff database. Following the protocol [15], we select 95 subjects as the training set, and the remaining 143 subjects are used for testing.

3.2 Experimental Settings

Images are cropped and resized to resolution of the 128×128 and converted to grayscale. We choose VGG-16 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. Experimental results in this paper are mean results of 20 random running.

3.2.1 Parameter Analysis of $m$ and $\lambda$

We take the CUFSF database as an example and analyze the parameters $m$ and $\lambda$. Figure 3 shows the recognition accuracy of our approach with different values of $m$ and $\lambda$ from 0.1 to 1.0 with step size 0.1. When the threshold values $m$ is between 0.3, 0.4 and 0.5, the experimental result is better and stable. For $\lambda$, we can achieve better and stable recognition accuracy, when it is set between 0.1 and 0.2. Finally, we set $m$ as 0.3, and $\lambda$ as 0.1 in our experiments. We can observe a similar phenomenon in the IIIT-D database.

3.3 Experimental Results

On the CUFSF database, we compare our approach with state-of-the-art methods, and the results are shown in Table 1. For synthesis-based methods of MWF [4] and Fast-RSLCR [3], 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 C-DoGOGH [1] and HOG-PCA [2], and common subspace-based methods including MvDA [5] and CMML [6], there exists some room to improve their recognition accuracy. For the deep learning methods of VGG [17] and Triplet-CNN [12], 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 1.32% (= 92.43% - 91.12%). The reasons for the improvement are:

![](Fig.3.jpg)

Fig. 3 The influences of $m$ and $\lambda$ on CUFSF database.

| Method category               | Methods     | Rank-1 recognition(%) |
|-------------------------------|-------------|------------------------|
| Synthesis-based Methods       | MWF [4]     | 74.00                  |
| Feature                       | Fast-RSLCR [3] | 75.94                 |
| Descriptor-based Methods      | C-DoGOGH [1] | 89.03                  |
| Common                        | HOG-PCA [2]  | 91.12                  |
| Subspace-based Methods        | MvDA [5]    | 55.00                  |
| Deep Learning Methods         | CMML [6]    | 80.00                  |
| Our approach                  | VGG [17]    | 48.82                  |
|                               | Triplet-CNN [12] | 86.32            |

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

![Fig.2](examples/face_sketch_photos.png)

(a) (b)

Fig. 2 Examples of face sketch-photos on (a) CUFSF database and (b) IIIT-D database. The first row shows the original photos and the second row shows the sketches.
Table 2 Average rank-1 recognition accuracy (%) on the IIIT-D database.

| Methods       | MCWLD [13] | RCN-10 [9] | CDL [16] | Triplet-CNN [12] | Ours |
|---------------|------------|------------|----------|------------------|------|
| Rank-1(%)     | 84.24      | 90.34      | 85.35    | 84.07            | **91.40** |

our approach makes an effort to enlarge the distance between inter-class samples and reduce the distance between intra-class samples.

Due to the substantial discrepancies in the IIIT-Delhi Viewed Sketch database, the sketch face recognition is much more difficult than on the CUFSF database. We compare our approach with the representative methods in Table 2. Our approach is compared with some state-of-the-art methods, including MCWLD [13], RCN-10 [9], CDL [16], and Triplet-CNN [12]. From the table, RCN-10 achieves good result, and our approach performs the best and improves accuracy by 1.06% (= 91.40% - 90.34%).

4. Conclusions

In this paper, we build a three channel CNN architecture and design a triplet-margin-center loss function to learn the discriminative feature representations. In our loss function, we reduce intra-class variations and enlarge the inter-class discrepancy. In addition, we propose a hard sample selection strategy to select hard samples, and avoid the slow convergence in training. Experimental results on two databases demonstrate the effectiveness and superiority of the proposed method.

Acknowledgments

The work described in this paper was fully supported by the National Natural Science Foundation of China (No. 61702280), Natural Science Foundation of Jiangsu Province (No. BK20170900), National Postdoctoral Program for Innovative Talents (No. BX20180146), China Postdoctoral Science Foundation (No. 2019M661901), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 2019K024), CCF-Tencent Open Fund WeBank Special Funding (No. CCF-WebankRAGR20190104), and Scientific Research Starting Foundation for Introduced Talents in NJUPT (NUPTSF, No. NY217009).

References

[1] S. Setumin and S.A. Suandi, “Cascaded static and dynamic local feature extractions for face sketch to photo matching,” IEEE Access, vol.7, pp.27135–27145, 2019.
[2] A. Radman and S.A. Suandi, “Robust face pseudo-sketch synthesis and recognition using morphological-arithmetic operations and HOG-PCA,” Multimedia Tools and Applications, vol.77, no.19, pp.25311–25332, Feb. 2018.
[3] N. Wan, X. Gao, and J. Li, “Random sampling for fast face sketch synthesis,” Pattern Recognit., vol.76, pp.215–227, April 2018.
[4] H. Zhou, Z. Kuang, and K.-Y.K. Wong, “Markov weight fields for face sketch synthesis,” IEEE Conference on Comput. Vis. Pattern Recognit., pp.1091–1097, 2012.
[5] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, “Multi-view discriminant analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol.38, no.1, pp.188–194, 2016.
[6] A. Mignon and F. Jurie, “CMML: A new metric learning approach for cross-modal matching,” IEEE Conference on Comput. Vis. Pattern Recognit., pp.1–14, 2012.
[7] C. Galea and R.A. Farrugia, “Forensic face photo-sketch recognition using a deep learning-based architecture,” IEEE Signal Process. Lett., vol.24, no.11, pp.1586–1590, Nov. 2017.
[8] P. Mittal, M. Vatsa, and R. Singh, “Composite sketch recognition via deep network-a transfer learning approach,” Int. Conf. Biometrics, pp.251–256, 2015.
[9] Z. Deng, X. Peng, and Y. Qiao, “Residual compensation networks for heterogeneous face recognition,” AAAI Conference on Artificial Intelligence, vol.33, no.1, pp.8239–8246, 2019.
[10] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” IEEE Conference on Comput. Vis. Pattern Recognit., pp.1335–1344, 2016.
[11] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, “Triplet-center loss for multi-view 3d object retrieval,” IEEE Conference on Comput. Vis. Pattern Recognit., pp.1945–1954, 2018.
[12] W. Wan, Y. Gao, and H.J. Lee, “Transfer deep feature learning for face sketch recognition,” Neural Computing and Applications 31, 9175-9184, 2019.
[13] W. Zhang, X. Wang, and X. Tang, “Coupled information-theoretic encoding for face photo-sketch recognition,” IEEE Conference on Comput. Vis. Pattern Recognit., pp.513–520, 2011.
[14] X. Wang and X. Tang, “Face photo-sketch synthesis and recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.11, pp.1955–1967, Nov. 2009.
[15] H.S. Bhatt, S. Bharadwaj, and M. Vatsa, “Memetic approach for matching sketches with digital face images,” IIITD-TR-2011-006, 2012.
[16] X. Wu, L. Song, R. He, and T. Tan, “Coupled deep learning for heterogeneous face recognition,” AAAI Conference on Artificial Intelligence, 2018, pp.1679–1686.
[17] O.M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” The British Machine Vision Conference, pp.1–12, 2015.