Orientation truncated centre learning for deep face recognition

Monica M.Y. Zhang, Yifang Xu✉ and Huaming Wu

Recently, centre loss that aiming to assist Softmax loss with the objectives of both inter-class dispension and intra-class compactness simultaneously, has achieved remarkable performance on convolutional neural network-based face recognition. However, its advantages highly rely on the centre feature assumption, which influences the capacity of the final obtained face features. Inspired by the centre loss approach, a novel Orientation Truncated Centre Learning is proposed, which takes advantage of an orientation truncated centre function to make the centre feature learning have more suitable orientation for deep face recognition. Three metrics are proposed to evaluate how discriminative are the distributions of the learned features for MNIST visualisation. Experimental results on several challenging benchmarks, including fine-grained labelled faces in the wild (FGLFW), labelled faces in the wild (LFW), YouTube faces (YTF), and benchmark of large-scale unconstrained face recognition (BLUFR), show that the proposed approach can easily generate more favourable results than several state-of-the-art competitors.

Introduction: Convolutional neural network (CNN) based face recognition has achieved significant performance. However, how to design better supervision signals for more discriminative face features is one of the most concerned issues. Commonly used loss functions include Softmax loss, Contrastive loss [1] and Triplet loss [2]. Softmax loss is effective for multi-class classification, but the learned face features are not discriminative sometimes. Contrastive loss and Triplet loss make it more discriminative by using information of feature pairs and triplets. However, the training procedure is not straightforward and the computation complexity will increase by selecting meaningful image pairs and triplets. Recently, centre loss approach [3], which is a simple and trainable method, has achieved great progress. However, its advantages highly rely on the centre feature assumption. Once the centre feature is not learned appropriately, the final face features may not represent the raw face images suitably. Particularly, the situation may be more serious when there exist a certain number of outliers. To this end, we propose a simple and efficient approach for more discriminative face features, called orientation truncated centre learning (OTCL) (see Fig. 1). Rather than defining the centre feature by averaging the features of the focused class in each iteration, OTCL learns the centre feature by using an orientation truncated function. The idea is to update the centre feature according to its nearest feature members, instead of using the full features to avoid the disturbance of some outliers. Thus, the centre feature can represent the features of the same class more efficiently to learn more suitable CNN models. Experimental results show the superiority of the proposed approach over two baselines and several state-of-the-art methods.

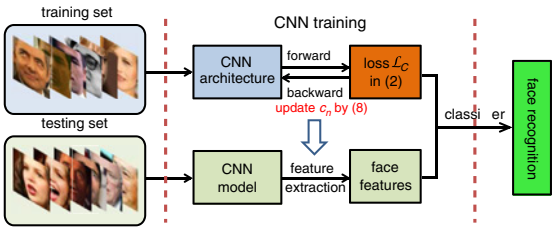

Fig. 1 *Framework of OTCL*

Proposed approach: Centre loss approach for CNN model learning is based on an optimisation objective, expressed as

$$\boldsymbol{\theta}^* = \min_{\boldsymbol{\theta}} \mathcal{L}_C(X, L, \boldsymbol{\theta}), \tag{1}$$

where $\mathcal{L}_C(X, L, \boldsymbol{\theta})$ is the joint supervision of Softmax loss $\mathcal{L}_S$ and centre loss $\mathcal{L}_c$, namely

$$\mathcal{L}_c(X, L, \boldsymbol{\theta}) = \mathcal{L}_S(X, L, \boldsymbol{\theta}) + \lambda \mathcal{L}_c(X, L, \boldsymbol{\theta}), \tag{2}$$

and $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is the training dataset, $L = \{l_1, l_2, \ldots, l_n\}$ is the corresponding label set, and $\boldsymbol{\theta}$ is the parameter set, λ is a hyper-parameter to balance the two losses. Here $\mathcal{L}_c$ is centre loss which is based on the distance of the feature $\boldsymbol{x}_m$ to its corresponding centre feature $\boldsymbol{c}_{l_m}$, formalised as

$$\mathcal{L}_C = \frac{1}{2M} \sum_{m=1}^{M} \left\| \boldsymbol{x}_m - \boldsymbol{c}_{l_m} \right\|^2, \tag{3}$$

where $\boldsymbol{c}_{l_m}$ is computed as the average of the features in the l_mth class, and M is the mini-batch size. However, when there exist many outliers for a focused class, the corresponding centre feature may not properly represent the class. As shown in Fig. 2*a*, many features are far away from their corresponding centres, which seem to have little connection with the centre feature updating.

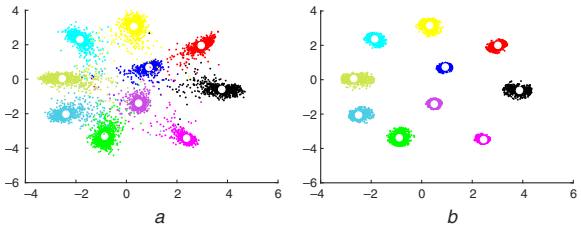

Fig. 2 *Centre features (white points) for different distributions of MNIST*
a Distribution of MNIST testing database by $\mathcal{L}_C$
b Distribution of MNIST testing database by $\mathcal{L}_{\text{OTC}}$

Intuitively, we can update the centre feature according to those nearest features around the centre feature, instead of using the full features. Suppose that there exist some nearest features $\{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_{N_i}}\}$ around the centre feature c_i, such that

$$\sum_{j=1}^{N_i} \left\| \boldsymbol{x}_{i_j} - \boldsymbol{c}_i \right\|^2 \approx f(i) = R \sum_{m=1}^{M} \mathbb{I}(l_m = i) \left\| \boldsymbol{x}_m - \boldsymbol{c}_{l_m} \right\|^2, \tag{4}$$

where N_i is the number of features in class i, $\mathbb{I}$ is the indicator function, and $R \in (0, 1)$. We want to find a suitable R to represent $\boldsymbol{c}_i$, and thus to avoid the disturbance of the outliers for centre feature updating, as shown in Fig. 2*b*.

For CNN training with N classes, considering all features in a mini-batch, then

$$\sum_{i=1}^{N} f(i) = R \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{I}(l_m = i) \left\| \boldsymbol{x}_m - \boldsymbol{c}_{l_m} \right\|^2 = R \sum_{m=1}^{M} d_m, \tag{5}$$

where $d_m = \| \boldsymbol{v}_m \|^2$ and $\boldsymbol{v}_m = \boldsymbol{x}_m - \boldsymbol{c}_{l_m}$, we aim to find a smallest $\hat{M}$ ($\hat{M} \leq M$) such that

$$\sum_{m=1}^{\hat{M}} d_{i_m} \geq R \sum_{m=1}^{M} d_m, \tag{6}$$

where $d_{i_1} \leq d_{i_2} \leq \cdots \leq d_{i_M}$. Then, we propose the orientation truncated centre (OTC) function

$$\mathcal{L}_{\text{OTC}} = \frac{1}{2M} \sum_{m=1}^{\hat{M}} d_{i_m} = \frac{1}{2M} \sum_{m=1}^{\hat{M}} \left\| \boldsymbol{x}_{i_m} - \boldsymbol{c}_{l_{i_m}} \right\|^2, \tag{7}$$

A truncated version of centre loss, to assist the centre feature updating to have a more suitable orientation for CNN feature extraction. Further, we update the centre feature by

$$\Delta \boldsymbol{c}_i = -\gamma \frac{\partial \mathcal{L}_{\text{OTC}}}{\partial \boldsymbol{c}_i} = \frac{\gamma}{M} \sum_{m=1}^{\hat{M}} \mathbb{I}(l_{i_m} = i) \boldsymbol{v}_{i_m}, \tag{8}$$

where γ is the centre feature learning rate.

In this way, we propose OTCL by changing the backward computation of centre loss approach by (8), without modifying the forward computation, which can be easily optimised by the standard stochastic gradient descent.

MNIST visualisation: We use LeNet++ [3] and MNIST database for feature visualisation. Three metrics are proposed to characterise the discrimination of the features: the average cosine distance between each sample and its corresponding centre feature (CD1), the average cosine distance between the centre features (CD2), the average cosine distance between each sample and its inter-class centre feature (CD3), where

$$\text{CD1} = \sum_{i=1}^{N} \sum_{j=1}^{N_i} \frac{1}{N_i N} \frac{\boldsymbol{c}_i^T \boldsymbol{x}_{ij}}{\|\boldsymbol{c}_i\| \|\boldsymbol{x}_{ij}\|}, \tag{9}$$

$$\text{CD2} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{I}(i \neq j) \frac{\boldsymbol{c}_i^T \boldsymbol{c}_j}{\|\boldsymbol{c}_i\| \|\boldsymbol{c}_j\|}, \tag{10}$$

$$\mathrm{CD3} = \frac{1}{\sum_i^N N_i(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N_i} \sum_{n=1}^{N} \mathbb{I}(n \neq i) \frac{\boldsymbol{c}_n^T \boldsymbol{x}_{ij}}{\|\boldsymbol{c}_n\| \|\boldsymbol{x}_{ij}\|}, \quad (11)$$

$\boldsymbol{c}_i$ is the centre feature for class i, $\boldsymbol{x}_{ij}$ is the feature in the class i, N_i is the number of features for class i, and N is the number of classes. By the above definitions, feature distribution with larger CD1, smaller CD2 and smaller CD3 is treated as more discriminative. As shown in Fig. 3, our proposed OTCL performs better than the centre loss approach.

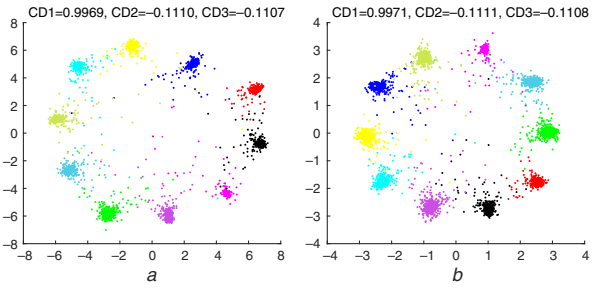

Fig. 3 *Final feature distributions corresponding to $\mathcal{L}_C$ and $\mathcal{L}_{OTC}$*
a Diameter of a class cluster is about 2
b Diameter of a class cluster is about 1

Experimental results: The proposed approach is used for face feature extraction without fine-tuning operations on CASIA-WebFace database [4] and ResNet-27 [3]. The initial learning rate is 0.1 and is divided by 10 at 30, 50 K iterations, until reaching the maximum iteration 60 K. We set $\lambda = 0.003$ and $\gamma = 0.5$ according to [3], and range R in [0.1, 0.2, …, 0.9].

Table 1: Comparing performance on FGLFW

Method	#Train (M)	FGLFW (%)
noisy Softmax [5]	0.5	94.50
human [6]	n/a	92.00
Deep convolutional maxout network (DCMN) [6]	0.5	91.00
Visual geometry group (VGG) [6, 7]	2.6	85.78
DeepFace [6, 8]	0.5	78.78
DeepID2 [1, 6]	0.2	78.25
Softmax	0.44	90.87
Softmax + Centre	0.44	94.28
OTCL-0.5	0.44	95.38
OTCL-0.7	0.44	95.45

Table 2: Comparing performance on LFW and YTF

Method	#Train (M)	LFW (%)	YTF (%)
SphereFace [9]	0.49	99.42	95.0
SphereFace	0.44	99.12	92.98
NormFace [10]	0.49	99.19	94.72
NormFace	0.44	98.63	93.26
Softmax + Centre [3]	0.7	99.28	94.9
Softmax + Centre	0.44	99.03	93.3
OTCL-0.5	0.44	99.17	93.94
OTCL-0.7	0.44	99.17	94.18

Table 3: Comparing performance on BLUFR protocol

Method	Verification (%)		Identification (%)	
	FAR = 0.1%	FAR = 1%	FAR = 1%	FAR = 10%
NormFace [10]	95.83	—	77.18	—
Softmax +Centre [3, 10]	93.35	—	67.86	—
LightenedCNN [11]	89.12	—	61.79	—
WebFaceCNN [4]	80.26	—	28.9	—
Softmax	82.22	93.5	56.81	73.3
Softmax + Centre	93.64	98.12	70.73	86.91
OTCL-0.5	94.15	98.15	75.29	88.73
OTCL-0.7	94.88	98.42	77.28	89.12

The performances of our best models OTCL-0.5 ($R = 0.5$) and OTCL-0.7 ($R = 0.7$) are reported on FGLFW in Table 1, LFW and YTF in Table 2, and BLUFR protocol in Table 3, respectively. Experimental results show that the proposed approach outperforms two baselines: Softmax, and Softmax + Centre. Specifically, our proposed approach achieves the first place on FGLFW and performs better than most of the compared methods on LFW, YTF, and BLUFR protocol. Note that it also surpasses NormFace and SphereFace with the same training data in Table 2. These all show the superiority of the proposed approach to characterise the centre features for more discriminative face features (Fig. 4).

Fig. 4 *Example images for MINST, LFW and YTF*

Conclusions: In this Letter, we propose a simple and more efficient algorithm for CNN-based face features learning, referred to as orientation truncated centre learning. By adopting an OTC function to restrict the clustering degree in a mini-batch for the centre feature definition, we make the centre feature represent the features of the same class more efficiently to learn CNN models. Feature visualisation with three metrics on real-world dataset Modified National Institute of Standards and Technology (MNIST) shows that the proposed approach can make the features more discriminative. Various evaluation implementations on face recognition tasks show that the proposed approach is effective and can easily generate more favourable results than the baseline centre loss approach and related state-of-the-art methods.

Submitted: *17 April 2018* E-first: *10 August 2018*
doi: 10.1049/el.2018.1326
One or more of the Figures in this Letter are available in colour online.

Monica M.Y. Zhang and Yifang Xu (*Center for Combinatorics, LPMC, Nankai University, No. 94, Weijin Road, Tianjin 300071, People's Republic of China*)

✉ E-mail: xyf@mail.nankai.edu.cn

Huaming Wu (*Center for Applied Mathematics, Tianjin University, No. 92, Weijin Road, Tianjin 300072, People's Republic of China*)

References

1 Sun, Y., Chen, Y.H., Wang, X.G., *et al.*: 'Deep learning face representation by joint identification-verification'. Proc. of NIPS, Montreal, QC, Canada, December 2014, pp. 1988–1996
2 Schroff, F., Kalenichenko, D., and Philbin, J.: 'Facenet: a unified embedding for face recognition and clustering'. Proc. of IEEE CVPR, Boston, MA, June 2015, pp. 815–823
3 Wen, Y.D., Zhang, K.P., Li, Z.F., *et al.*: 'A discriminative feature learning approach for deep face recognition'. Proc. of ECCV, Amsterdam, The Netherlands, October 2016, pp. 499–515
4 Yi, D., Lei, Z., Liao, S.C., *et al.*: 'Learning face representation from scratch'. *arXiv preprint arXiv:1411.7923*, 2014
5 Chen, B.H., Deng, W.H., and Du, J.P.: 'Noisy softmax: improving the generalization ability of dcnn via postponing the early softmax saturation'. Proc. of IEEE CVPR, Honolulu, HI, USA, July 2017
6 Deng, W.H., Hu, J.N., Zhang, N.H., *et al.*: 'Fine-grained face verification: FGLFW database, baselines, and human-dcmn partnership', *Pattern Recognit.*, 2017, **66**, pp. 63–73
7 Parkhi, O., Vedaldi, A., Zisserman, A., *et al.*: 'Deep face recognition', *Proc. BMVC*, 2015, **1**, p. 6
8 Taigman, Y., Yang, M., Ranzato, M., *et al.*: 'Deepface: closing the gap to human-level performance in face verification'. Proc. of IEEE CVPR, Columbus, OH, USA, June 2014, pp. 1701–1708
9 Liu, W.Y., Wen, Y.D., Yu, Z.D., *et al.*: 'Sphereface: deep hypersphere embedding for face recognition'. Proc. of IEEE CVPR, Honolulu, HI, USA, July 2017, pp. 212–220
10 Wang, F., Xiang, X., Cheng, J., *et al.*: 'Normface: L2 hypersphere embedding for face verification'. Proc. of ACM Multimedia, Mountainview, CA, USA, October 2017
11 Wu, X., He, R., and Sun, Z.N.: 'A lightened cnn for deep face representation', *arXiv preprint, arXiv*:1511.02683, 2015