OUTLIER-SUPPRESSED TRIPLET LOSS WITH ADAPTIVE CLASS-AWARE MARGINS FOR FACIAL EXPRESSION RECOGNITION

Yi Tian[1,2], *Zhiwei Wen*[1,2], *Weicheng Xie*[1,2*], *Xi Zhang*[1,2], *Linlin Shen*[1,2],*Jinming Duan*[3]

[1]School of Computer Science & Software Engineering , Shenzhen University, China
[2]Guangdong Key Laboratory of Intelligent Information Processing
[3]School of Computer Science, University of Birmingham, UK

ABSTRACT

Triplet loss has been proposed to increase the inter-class distance and decrease the intra-class distance for various tasks of image recognition. However, for facial expression recognition (FER) problem, the fixed margin parameter does not fit the diversity of scales between different expressions. Meanwhile, the strategy of selecting the hardest triplets can introduce noisy guidance information since various persons may present significantly different expressions. In this work, we propose a new triplet loss based on class-aware margins and outlier-suppressed triplet for FER, where each pair of expressions, e.g. ’happy’ and ’fear’, is assigned with an adaptive margin parameter and the abnormal hard triplets are discarded according to the feature distance distribution. Experimental results of the proposed triplet loss on the FER2013 and CK+ expression databases show that the proposed network achieves much better accuracy than the original triplet loss and the network without using the proposed strategies, and competitive performance compared with the state-of-the-art algorithms.

Index Terms— triplet loss, class-aware margin, outlier suppression, facial expression recognition

1. INTRODUCTION

While the softmax loss reflecting the recognition accuracy treats samples uniformly without considering specific prior information, a few deep metric learning algorithms have been proposed to embed specific constraint into the loss function of deep networks to improve the discriminative capability of the learned features. For example, the Center loss [1], feature loss [2] and adaptive deep metric learning [3] have been applied for object verification and re-identification.

Generally, deep metrics are proposed to increase inter-class distance and decrease intra-class distance. The triplet loss [4] is most intuitive among these deep metrics, whose variants have achieved competitive performance for various tasks. Motivated from the triplet loss, Chen et al. [5] proposed the quadruplet loss to further enlarge inter-class distance. Among the variants of the triplet loss, the triplet selection and the margin parameter adjustment get a lot of attention.

For the selection of triplets, Song et al. [6] selected the hardest sample pair in the batch training samples. Xiao et al. [7] proposed to select the most dissimilar sample pair with the same identity and the most similar sample pair with different identities in the batch training samples. Yuan et al. [8] divided negative samples into easy, semi-hard and hard samples and devised a cascaded network to handle the samples differently. Hermans et al. [9] selected the hardest triplet according to the distances from the anchor sample for person re-identification. Sohn [10] proposed a multi-class N-pair loss to make the negative examples interact with each other to reduce the influence of slow convergence of the original triplet loss. Wu et al. [11] proposed to select the hard triplets with uniform distance distribution according to the distance probability. Yu et al. [12] proposed an adapted triplet loss to learn the feature embedding by minimizing the distribution shift on the selected triplets. These algorithms selected the hard triplets under the assumption that the selected samples are not significantly different from those from the same class. However, for tasks like facial expression recognition (FER), one person may have an expression significantly different from others, learning from which may introduce misguidance information and result in poor generalization ability. In this work, we use the feature distance distribution to detect the abnormal samples and exclude them from the hard triplet selection.

Considering the margin parameter setting, Hadsell et al. [13] introduced a margin parameter in the contrastive loss to select the difficult training samples for dimensionality reduction. While the margin parameter determines the range of the negative and positive samples selection for triplet loss learning, its self-adaptive update was often employed. Wu et al [14] proposed a heuristic strategy to update the margin pa-

[*]The work was supported by Natural Science Foundation of China under grants 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 grants no. 2018063. Corresponding author: Dr. Weicheng Xie, wcxie@szu.edu.cn

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

rameter with large intra-class cosine distances, while Li et al. [15] used the inter-class distances of the projected image features, Zakharov et al. [16] used the similarity of the predicted poses, and Wang et al. [17] and Chen et al. [5] used the intra-class and inter-class distances for the self-adaption of margins. Motivated from the study [5], we use distances between the class centers to approximate the variations of the inter-class distance, which not only decreases the time complexity of inter-class distance computation, but also improves the robustness by using more global information since centers are updated according to all the visited training samples, rather than the samples from current training batch.

Though margin parameter can be made self-adaptive, a unique margin for all the triplets may neglect the deform intensity inconsistency among different sample classes. For the FER problem, the variation between 'fear' and 'sad' expressions is significantly smaller than that between 'fear' and 'happy'. Thus, Kyperountas et al. [18] and Xie et al. [19] proposed the pairwise and triplet-wise FER to take into account the characteristic of each expression group. Triplet loss was studied on FER [20] and Liu et al. [3] proposed the (N+M)-tuplet loss to allow interaction among multiple positive samples. However, customizing the margin parameter for each expression pair based on its scale specificity is rarely studied.

In this work, a new triplet loss with class-aware margins is proposed for FER, which makes the following contributions.

- The deform intensity inconsistency among expressions is taken into account by introducing a specific margin for each expression pair, and the margins are self-adaptively updated;
- Noisy hard triplets are reduced according to feature distance distribution to reduce the influence of abnormal expression samples;
- The proposed algorithm achieves competitive performance on two public expression databases, compared with the original triplet loss and the state-of-the-art approaches.

2. THE PROPOSED ALGORITHM

For the face alignment of image preprocessing, the five key points are firstly located on the eyes, nose and the mouth tips [21]. Then the database is augmented by cropping different regions. Each expression image I is then normalized in gray level, mirrored and scaled to the size 227×227 for the training.

Before the introduction of the proposed loss, the original triplet loss is presented as follows

$$\mathcal{L}_t^{ori} = \frac{1}{2}\sum_{x_a}[d(f(x_a), f(x_p))^2 - d(f(x_a), f(x_n))^2 + \alpha]_+, \tag{1}$$

where α is the margin determining the range for triplet selection, x_p and x_n, i.e. the positive and negative samples selected randomly from two different classes, together with the anchor sample $x_a (x_p \neq x_a)$ from the same class as x_p, compose a triplet; $d(f(x_n), f(x_a)) = ||f(x_n) - f(x_a)||_2$ is the L_2-norm distance; $f(x_a)$ is the embedded feature vector, i.e. network output of the fully connected (FC) layer of the anchor (x_a) sample. x_a and $f(x_a)$ determine the gradient direction for back propagation; $[\cdot]_+$ is the hinge function and formulated as $max(\cdot, 0)$.

2.1. Self-adaptive class-aware margins

In order to consider the scale specificity of each expression pair, a margin is assigned to each expression pair ($\#class(\#class - 1)/2$ margins), i.e. the margin is assumed to be dependent on the expression classes of the triplet to consider the characteristics of each expression. The class-aware triplet loss is formulated as follows

$$\mathcal{L}_t = \frac{1}{2}\sum_{x_a}[d(f(x_a), f(x_p))^2 - d(f(x_a), f(x_n))^2 + \alpha_m]_+, \tag{2}$$

where α_m is the margin associated with the m-th triplet (x_a, x_p, x_n) and reflects the characteristics of each expression triplet. Since the triplet (x_a, x_p, x_n) is closely related to the labels of (x_a, x_n), we devise expression pair-aware margins for the triplet loss.

Since large margin parameter encourages more hard triplets, an online margin updating method is proposed to use the updated centers of the expression classes as follows

$$\begin{cases} nr = \frac{1}{2}min(\frac{1000}{NumIter}, 1), \\ dc_{i,j} = ||c_i - c_j||_2, \\ \alpha_m^{new} = [dc_{i,j} - \frac{1}{N_a}\sum_{a,p} d(f(x_a), f(x_p))^2]_+, \\ \alpha_m = (1 - nr) \cdot \alpha_m^{old}\frac{\gamma^{new}}{\gamma^{old}} + nr \cdot \alpha_m^{new}. \end{cases} \tag{3}$$

where γ^{new}, γ^{old} are the L_2-norms of the embedded features of the last and current iterations, i. e. $||f(x_a)||_2 = ||f(x_p)||_2 = ||f(x_n)||_2 = \gamma$ [22]; N_a is the number of the intra-class sample pairs; the weight nr is introduced to use the preceding information for margin update; the class centers $\{c_i\}$ are updated during the back propagation of the center loss $\mathcal{L}_C$ [1] as follows

$$\mathcal{L}_C = \frac{1}{2}\sum_i d(f(x_i), c_{y_i})^2, \tag{4}$$

where y_i is the expression label of x_i, c_{y_i} is the center vector of the y_i-th class features $\{f(x_i)\}$.

2.2. Outlier-suppressed hard triplet selection

According to the study [7], the hardest positive and negative samples are selected as follows

$$\begin{cases} x_p^* = arg\max_{x_p} d(f(x_a), f(x_p))^2, \\ x_n^* = arg\min_{x_n} d(f(x_a), f(x_n))^2, \end{cases} \tag{5}$$

However, learning from the hardest triplet may misguide the network training due to possible abnormal samples. In this work, we detect abnormal hard triplets according to feature distance distribution and discard them in advance.

According to the studies [23, 11], the random variable of feature distance d (Fig. 1(a)) approximately obeys the following normal distribution

$$d \sim \mathcal{N}(\sqrt{2}\gamma, \frac{\gamma}{\sqrt{2n}}), \tag{6}$$

where γ is the L_2-norm of the embedded feature; n is the feature dimension; $\sqrt{2}\gamma$ and $\frac{\gamma}{\sqrt{2n}}$ are the mean and standard variance of the normal distribution.

We assume a selected positive sample x_p is normal if the distance $d(f(x_a), f(x_p))$ falls in the acceptance region of a null hypothesis H_0 under a significance level τ_p, while abnormal if this distance lies in the corresponding rejection region, i.e. the corresponding alternative hypothesis H_1 is accepted. The null and alternative hypotheses are presented as follows

$$\begin{cases} H_0 : \{\mu_{d(f(x_a),f(x_p))} \leq \sqrt{2}\gamma\}, \\ H_1 : \{\mu_{d(f(x_a),f(x_p))} > \sqrt{2}\gamma\}. \end{cases} \tag{7}$$

where $\mu_{d(f(x_a),f(x_p))}$ denotes the mean of the random variable $d(f(x_a), f(x_p))$. Thus, the triplet (x_a, x_p, x_n) is discarded when $d(f(x_a), f(x_p))$ or $d(f(x_a), f(x_n))$ lies in the corresponding rejection region, i.e. one of the following rejection conditions is satisfied

$$\begin{cases} d(f(x_a), f(x_p)) \leq \sqrt{2}\gamma + \frac{\gamma}{\sqrt{2n}} F^{-1}(1 - \tau_p), \\ d(f(x_a), f(x_n)) \geq \sqrt{2}\gamma + \frac{\gamma}{\sqrt{2n}} F^{-1}(\tau_n), \end{cases} \tag{8}$$

where $F^{-1}(1 - \tau_p)$ is the inverse of the cumulative probability distribution of the normal distribution in equation (6) with probability being $1 - \tau_p$, i.e. $P_F\{d \leq F^{-1}(1 - \tau_p)\} = 1 - \tau_p$; τ_p, τ_n are the significance levels of the positive and negative samples, respectively. An example distance variable and the rejection regions for triplet selection are presented in Fig. 1.

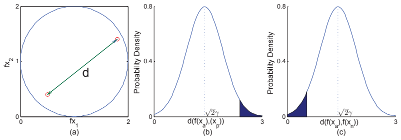

Fig. 1: (a) Random distance d with $n = 2$, where $f(x) = (fx_1, fx_2)$. (b),(c) The feature distance distribution and the rejection regions (blue solid regions) for hard positive and negative sample selection with significance levels of $\tau_p = 0.025$ and $\tau_n = 0.05$, respectively.

Compared with equation (2), the proposed outlier-suppressed method in equation (8) provides additional upper and lower bounds, which can decrease the influence of noisy samples.

2.3. Network training

The residual network (ResNet) [24] with slight modification, i.e. the number of the last FC output changed to #class, is used for the training and evaluation. The self-adaptive normalization layer [22] is added after the last but one FC layer, i.e. the L_2-norm of the FC output vector $f(x)$ is normalized to a value γ with a self-adaptive mode. To fully make use of the already trained models, the fine tuning of an available face recognition model is employed.

For the network training, the softmax loss $\mathcal{L}_S$ is used to boost the discriminative ability in addition to the center and proposed triplet losses. The final loss is then formulated as follows

$$\mathcal{L} = \mathcal{L}_S + \lambda_C \mathcal{L}_C + \lambda_t \mathcal{L}_t. \tag{9}$$

where $\mathcal{L}_C$ and $\mathcal{L}_t$ are the center and triplet loss presented in equations (4) and (2); λ_C and λ_t are the regularization coefficients. For the network SGD optimization, the gradient of $\mathcal{L}$ w.r.t. each variable is calculated for back propagation.

Since the centers are not stable in the preliminary iterations, to reduce the influence of such instability for the margin update (equation (3)), a scale factor ρ before the loss weight of λ_t (2) is introduced, and gradually increased from 0 to its maximum as follows

$$\rho(\#iter) = \frac{1}{1 + 10e^{-\frac{\#iter}{3000}}}, \tag{10}$$

where $\#iter$ is the number of algorithm iterations. For the recognition of each testing sample, majority voting based on the probabilities of augmented face regions is employed.

3. EXPERIMENTAL RESULTS

We perform the experiments using four-kernel Nvidia TITAN GPU Card and CAFFE package. The parameter settings of the proposed algorithm are presented in Table 1.

The proposed algorithm is tested on the expression databases of the Extended Cohn-Kanade Dataset (CK+) [25] and FER2013 database [26], whose examples are presented in Fig. 2. These databases are categorized to six basic and the neutral expressions, i.e. angry (An), disgust (Di), fear (Fe), happy (Ha), sad (Sa), surprise (Su) and neutral (Ne). The CK+ database consists of 593 expression sequences from 123 subjects, and 1033 expression images, i.e., the neutral and three non-neutral images sampled from each expression sequence are used for testing. The person-independent strategy with ten-fold cross validation is employed for testing and comparison. The FER2013 database [26] 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 both the validation and testing sets consist of 3,589 samples.

To test the performance of the proposed triplet loss, the ResNet network is trained with different loss strategies, i.e. the class-aware margins, the outlier-suppressed triplet selection and their combination, then their performances on the FER2013 database are presented in Table 2.

Table 1: The parameter setting of the proposed algorithm.

Parameter	Value	Parameter	Value	Parameter	Value
λ_C	8e-3	λ_t	1e-5	Learning rate	5e-3
Batch size	60	τ_p	0.025	τ_n	0.05

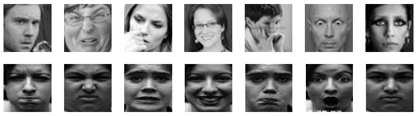

Fig. 2: Example images of FER2013 and CK+. The columns represent expressions of An, Di, Fe, Ha, Sa, Su and Ne, respectively.

Table 2: The performances (%) with different loss settings for the FER2013 database.

Method	Recog. Rate (%)
Baseline (Softmax Only)	68.91
Original hard triplet [7] with center loss	71.41
Our 21 margins	72.14
Our hard triplet selection with outlier suppression	71.86
Our 21 margins+Outlier suppression	**72.64**

In Table 2, the performance with the softmax loss SM is listed as the benchmark. Table 2 shows that the proposed algorithm with both 21 margins and outlier suppression achieves an improvement of 3.73% over the benchmark setting, while class-aware margins and outlier suppression achieved improvements of 2.95% and 3.23% over the baseline on the FER2013 database.

Fig. 3 demonstrates three example outliers detected by the proposed outlier suppression approach during positive sample selection, whose anchor-positive distances lie in the rejection region of the distance distribution. One can observe that the expressions labeled with 'angry', 'neutral' and 'fear' in Fig. 3 can be easily confused with 'surprise', 'sad' and 'sad',respectively. Thus, the exclusion of these confusing expressions can reduce misguidance information during network training and help the network to generalize well to other expressions, which illustrates the usefulness of the proposed outlier suppression approach.

To compare the performance of the proposed algorithm with other algorithms, Table 3 lists the recognition rates of the proposed approach, state-of-the-art algorithms and the baselines on the FER2013 and CK+ databases. For the FER2013 database, our algorithm achieved as high as 72.64% accuracy, which is even 1.44% higher than that of the challenge winner [27], i.e. 71.2%. For the CK+ database, Table 3 shows that the proposed algorithm achieves a recognition rate of 97.11%, which ranked the third among five state-of-the-art algorithms and an improvement of 2.41% is achieved over the baseline of Softmax loss.

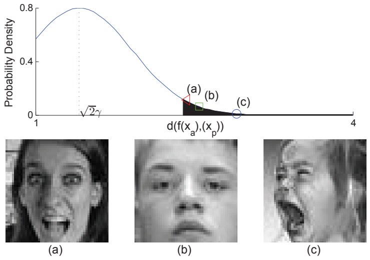

Fig. 3: Three outliers detected with the proposed outlier suppression approach during positive sample selection.

Table 3: Comparison of different algorithms on the FER2013 (FER.) and CK+ databases.

Data.	Algorithm	Subjects	Protocol	Recog. rate (%)
FER.	DNN with SVM [27]	-	-	71.2
	Feature loss [2]	-	-	61.86
	Network with Softmax loss	-	-	68.91
	Ours	-	-	**72.64**
CK+	DNN [28]	106	5-fold	93.2
	Patch weighting [4]	106	10-fold	94.09
	Triplet-wise feature optimization [19]	106	10-fold	94.09
	De-expression network [29]	118	10-fold	97.3
	Feature loss [2]	106	10-fold	**97.35**
	Network with Softmax loss	106	10-fold	94.7
	Ours	106	10-fold	97.11

4. CONCLUSION

This work proposed a class-aware and outlier-suppressed triplet loss for facial expression recognition (FER). The class-aware margins are introduced to address the deform intensity inconsistency of each expression pair. Meanwhile, during the triplet selection, the abnormal hard triplets are excluded according to feature distance distribution to reduce the influences of abnormal expressions. The experimental results on CK+ and FER2013 databases show the advantages of the proposed algorithm over the original triplet loss and the related state-of-the-art algorithms. In future, samples excluded with hard-triplet selection will be further classified and handled differently.

5. REFERENCES

[1] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *ECCV*, 2016, pp. 499-515.

[2] G. Zeng, J. Zhou, J. Xi, W. Xie, and L. Shen, “Hand-crafted feature guided deep learning for facial expression recognition,” in *FGR*, 2018, pp. 423-430.

[3] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, “Adaptive deep metric learning for identity-aware facial expression recognition,” in *CVPRW*, 2017, pp. 522-531.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015, pp. 815-823.

[5] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: a deep quadruplet network for person re-identification,” in *CVPR*, 2017, pp. 1320-1329.

[6] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, “Deep metric learning via lifted structured feature embedding,” in *CVPR*, 2016, pp. 4004-4012.

[7] Q. Xiao, H. Luo, and C. Zhang, “Margin sample mining loss: a deep learning based method for person re-identification,” in *arXiv:1710.00478*, 2017.

[8] Y. Yuan, K. Yang, and C. Zhang, “Hard-aware deeply cascaded embedding,” in *ICCV*, 2017, pp. 814-823.

[9] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” in *arXiv preprint arXiv:1703.07737*, 2017.

[10] K. Sohn, “Improved deep metric learning with multi-class N-pair loss objective,” in *NIPS*, 2016, pp. 1857-1865.

[11] C. Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, “Sampling matters in deep embedding learning,” in *ICCV*, 2017, pp. 2859-2867.

[12] B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao, “Correcting the triplet selection bias for triplet loss,” in *ECCV*, 2018, pp. 1-17.

[13] R. Hadsell, S. Chopra, and Y. Lecun, “Dimensionality reduction by learning an invariant mapping,” in *CVPR*, 2006, pp. 1735-1742.

[14] B. Wu, H. Wu, and M. M. Y. Zhang, “Scalable angular discriminative deep metric learning for face recognition,” in *arXiv:1804.10899v2*, 2018.

[15] Y. Li, Z. Jia, J. Zhang, K. Huang, and T. Tan, “Deep semantic structural constraints for zero-shot learning,” in *AAAI*, 2018.

[16] S. Zakharov, W. Kehl, B. Planche, A. Hutter, and S. Ilic, “3D object instance recognition and pose estimation using triplet loss with dynamic margin,” in *IROS*, 2017, pp. 552-559.

[17] J. Wang, S. Zhou, J. Wang, and Q. Hou, “Deep ranking model by large adaptive margin learning for person re-identification,” *PR*, vol. 74, no., pp. 241-252, 2017.

[18] M. Kyperountas, A. Tefas, and I. Pitas, “Salient feature and reliable classifier selection for facial expression classification,” *PR*, vol. 43, no. 3, pp. 972-986, 2010.

[19] W. Xie, L. Shen, Y. Meng, and Z. Lai, “Active AU based patch weighting for facial expression recognition,” *Sensors*, vol. 17, no. 2, pp. 275, 2017.

[20] J. Wang and C. Yuan, “Facial expression recognition with multi-scale convolution neural network,” in *PCM*, 2016, pp. 376-385.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.

[22] R. Ranjan, C. D. Castillo, and R. Chellappa, “L2-constrained softmax loss for discriminative face verification,” in *arXiv:1703.09507*, 2017.

[23] The sphere game in n dimensions, http://faculty.madisoncollege.edu /alehnen/sphere/hypers.htm, Accessed: 2018-09-27.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770-778.

[25] T. Kanade, Y. Tian, and J. F. Cohn, “Comprehensive database for facial expression analysis,” in *FGR*, 2000, pp. 46.

[26] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *ICME*, 2005, pp. 5.

[27] Y. Tang, “Deep learning using linear support vector machines,” in *ICML*, 2013.

[28] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in *WACV*, 2016, pp. 1-10.

[29] H. Yang, U. Ciftci, and L. Yin, “Facial expression recognition by de-expression residue learning,” in *CVPR*, 2018.