Attention-Set based Metric Learning for Video Face Recognition

Yibo Hu*†‡, Xiang Wu*† and Ran He*†‡
*Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China
†National Laboratory of Pattern Recognition, CASIA, Beijing, China
‡University of Chinese Academy of Sciences, Beijing, China
Email: yibo.hu@cripac.ia.ac.cn, alfredxiangwu@gmail.com, rhe@nlpr.ia.ac.cn

***Abstract*—Face recognition has made great progress with the development of deep learning. However, video face recognition (VFR) is still an ongoing task due to various illumination, low-resolution, pose variations and motion blur. In this paper, we propose a novel Attention-Set based Metric Learning (ASML) method for VFR. It is a promising and generalized extension of Maximum Mean Discrepancy with Memory Attention Weighting inspired by Neural Turing Machine. ASML can be naturally integrated into Convolutional Neural Networks, resulting in an end-to-end learning scheme. Our method achieves state-of-the-art performance for the task of video face recognition on three widely used benchmarks including YouTubeFace, YouTube Celebrities and Celebrity-1000.**

***Keywords*-video face recognition; metric learning; memory attention weighting;**

I. Introduction

Video face recognition (VFR) has attracted a significant attention in computer vision community in recent years [15], [35], [33], [19], [22], [5], [2], [9], [10], [29], [7]. In contrast to conventional face recognition where each gallery or probe instance refers to a single image, video face recognition aims to identify and verify face videos. Moreover, each video can be treated as an image set of faces without taking temporal information into account. Generally, there three core issues in VFR: how to alleviate sample biases and noise within a video or an image set, how to construct appropriate face representations, and how to define a suitable distance metric for calculating the similarity between these representations.

Recently, with the prominent success of Convolutional Neural Network (CNN) in various vision applications, some works [21], [33], [30], [19], [27], [24], [25] employ CNNs to automatically learn the mapping from the input face images to a discriminative embedding with supervised signals. The superiority of CNN is that the network can learn the face representations and the discriminative embeddings jointly in an end-to-end manner. Current CNN based face recognition approaches [21], [30], [19], [27], [24], [25] are originally designed for single face images rather than videos. On the one hand, due to the low-resolution and motion blur in videos, the performance of these approaches for VFR will decrease. On the other hand, these methods extract features for each image in a video and then simply aggregate them as a representation or fuse the matching results across all pairs of images between two videos. However, in this way, it doesn't consider the correlations of images in videos. Yang et al. [33] design a novel attention-based feature aggregation strategy named Neural Aggregation Network (NAN). It adaptively aggregates the face features in a video by advocating high-quality face images and suppressing low-quality ones. The existing CNN based methods for VFR either take shuffled single images or randomly selected pairs or hard sampled triplets as training instances, which can not make full use of rich information of face videos (or face image sets) in the CNN training stage.

In this paper, we propose an Attention-Set based Metric Learning (ASML) approach for video face recognition. It addresses the above challenges and can be integrated into general CNN frameworks seamlessly with end-to-end trainable parameters. ASML is a promising and generalized extension of Maximum Mean Discrepancy [20] with memory attention weighting. Each training instance is consisted of three parts: an anchor image set, a positive image set and a negative image set. Based on these triple-set instances, a siamese CNN with three branches is trained by supervision of ASML. ASML has powerful abilities to drive the CNN to learn more discriminative and set-aware face representations. Memory attention weighting is employed to correct sample biases in the triple-set instances. This policy can effectively maintain and focus on beneficial information while suppress and discard noisy information in network training. The proposed method is evaluated on three widely used datasets including YouTubeFace [31], YouTube Celebrities [11] and Celebrity-1000 [14] and achieves the best performance compared to state-of-the-art methods. The basic framework of our approach is illustrated in Figure 1.

Note that, the proposed ASML has huge distinctions with NAN. NAN focuses on attentively aggregating face features over a face video to obtain a single feature. But ASML concentrates on learning representative and discriminative features based on attention mechanism and just uses average aggregation to fuse these features. From this perspective, ASML and NAN are complementary and can be combined with each other seamlessly, because they are two independent parts.

II. Our Approach

In this section, we first introduce ASML, and then embed it into a general CNN framework via a simplified Neural Turing Machine [6]. Finally, we depict the whole network architecture.

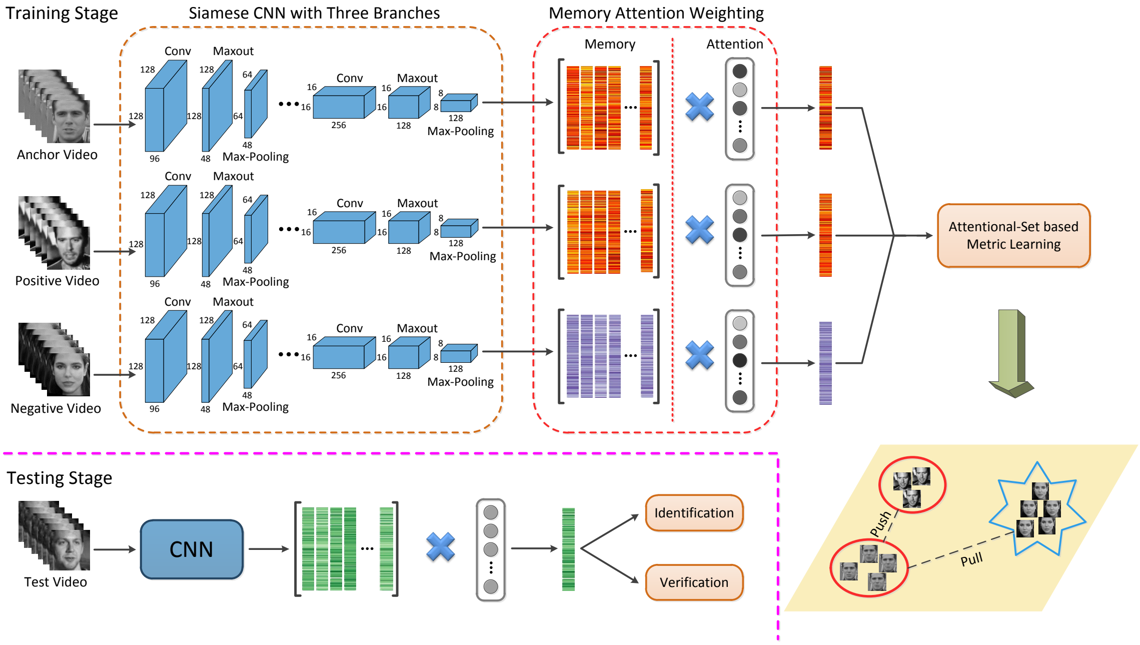

Figure 1. An illustration of our proposed approach. Different colors in attention blocks indicate different significance of the corresponding face features.

A. Attention-Set based Metric Learning

Generally, the key component in video face recognition is image set representations, more specifically is the image features [15], [8], [35]. Since the images from the same sets are generally similar, we expect they have similar (or same) feature distributions. On the other hand, the images from the different sets are always divergent and should have disparate feature distributions. ASML has the ability to learn such kind of features. It is a promising and generalized extension of Maximum Mean Discrepancy with memory attention weighting.

Maximum Mean Discrepancy (MMD) [20] was primitively proposed to measure the distance between two distributions based on the mean features of data sets in a Reproducing Kernel Hilbert Space (RKHS). Let p and q be distributions defined on a domain $\mathcal{X}$. Assuming two data sets $X=\{x_1, ..., x_n\}$ and Y $=\{y_1, ..., y_m\}$ are drawn i.i.d. from p and q respectively, the MMD criterion determines whether $p = q$ or $p \neq q$ in RKHS.

Definition 1: (from [20]) Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to R$ and let p, q, X, Y be defined as above. The MMD and its empirical estimate as:

$$\mathrm{MMD}\left[\mathcal{F}, p, q\right] := \sup_{f\in\mathcal{F}} \left(E_{x\sim p}\left[f\left(x\right)\right] - E_{y\sim q}\left[f\left(y\right)\right]\right) \quad (1)$$

$$\mathrm{MMD}\left[\mathcal{F}, X, Y\right] := \sup_{f\in\mathcal{F}} \left(\tfrac{1}{n}\sum_{i=1}^{n} f(x_i) - \tfrac{1}{m}\sum_{i=1}^{m} f(y_i)\right) \quad (2)$$

When $\mathcal{F}$ is a unit ball in RKHS which is defined on compact metric space $\mathcal{X}$, the equation $\mathrm{MMD}\left[\mathcal{F}, X, Y\right] = 0$ satisfies if and only if $p = q$ [20]. Intuitively, the smaller MMD the more correlative between distributions of X and Y, thus the features of the two sets are similar. Accordingly, the larger MMD the more discrepant between the distributions, and the features are disparate. Taken in this sense, we propose the following Mean-Set based Metric Learning (MSML),

$$\begin{aligned}\mathrm{MSML}(X, Y, Z) &= \|E_{x\sim p}[f(x)] - E_{y\sim p}[f(y)]\|_2 \\ &+[\alpha - \|E_{x\sim p}[f(x)] - E_{z\sim q}[f(z)]\|_2]_+\end{aligned} \quad (3)$$

where $[\,\cdot\,]_+$ indicates $max(\cdot, 0)$ and α is a constant margin. X, Y, Z represent image sets. Among them, X and Y are drawn from the same class, but different from Z. Minishing the quantity of MSML, one can increase the correlation and discrepancy between the similar and dissimilar sets respectively.

Although MMD can be applied to arbitrary dimensions and domains, it doesn't take into consideration the sample biases and outliers (or noise) in the sets, which is ubiquitous in real world applications. Consequently, it is inferior to VFR. To address it, from the proposition 1 in [4], we obtain that the following optimization problem is convex and has the unique solution $\omega(x) = p(x)/\hat{p}(x)$:

$$\begin{aligned}&\underset{\omega(x)\geq 0}{minimize}\ \|E_p[f(x)] - E_{\hat{p}}[\omega(x)f(x)]\|_2 \\ &s.t.\ E_{\hat{p}}[\omega(x)] = 1\end{aligned} \quad (4)$$

where $\hat{p}$ is a distribution whose support corresponds with p. Obviously, it is reasonable to believe that $\omega(x)$ can correct sample biases and remove outliers (or reduce noise), if it is well selected. Introducing $\omega(x)$ as a correction term of sample biases into MMD, we have Rectified Mean

Discrepancy (RMD) as follows:

$$\begin{aligned} \mathrm{RMD}\left(X,Y\right) &= \left\|E_{\omega(x)}[f(x)] - E_{\omega(y)}[f(y)]\right\|_2 \\ &= \| \sum_{i=1}^{n} \omega(x_i) f(x_i) - \sum_{j=1}^{m} \omega(y_j) f(y_j) \|_2 \\ s.t.\ &\sum_{i=1}^{n} \omega(x_i) = 1,\ \sum_{j=1}^{m} \omega(y_j) = 1 \end{aligned} \tag{5}$$

Combining Eq.(3) and Eq.(5), we couple Mean-Set based Metric Learning with the rectified term and obtain a superior metric learning method named Attention-Set based Metric Learning (ASML) for video face recognition,

$$\begin{aligned} \mathrm{ASML}(X,Y,Z) &= \left\|E_{\omega(x)}[f(x)] - E_{\omega(y)}[f(y)]\right\|_2 \\ &+ [\ \alpha - \left\|E_{\omega(x)}[f(x)] - E_{\omega(z)}[f(z)]\right\|_2\ ]_+ \\ s.t.\ &\sum \omega(x) = 1,\ \sum \omega(y) = 1,\ \sum \omega(z) = 1 \end{aligned} \tag{6}$$

Obviously, it has the same characteristics with MSML, pushing the distributions of same sets closely and pulling the distributions of different sets far apart simultaneously. Additionally, the rectified term is equipped in it to correct biases and reduce noise.

B. Memory Attention Weighting

Weights in ASML indicate the significance of images in a set. If a face image is frontal, well-illuminated and legible, a high weight should be assigned. Otherwise, if it is lateral, bad-illuminated and blurred, a low weight is applied. From this perspective, two primary principles must be considered for weighting: First, the weighting method is easily integrated in basic CNN frameworks and its parameters are end-to-end trainable by means of supervised manner. Second, the weights should be global-content-based and set-aware, since we construct a training instance as a special image set to make better use of set information.

The memory attention mechanism [33], [23], [1], [6] is suitable for the above requirements. The core concept is to couple neural models with external memory and to be interacted by attentional addressing processes. In ASML, we regard the weighting as a Neural Turing Machine (NTM) [6], where the face feature sets are treated as memory and the weights are deemed to address to read from and write to memory attentively.

A NTM involves two basic components: a read head and a write head. They emit normalised outputs over memory locations to define the degree to which the head reads or writes. Therefore, a head can attend sharply at a single location or weakly ate multiple locations. In our memory attention weighting method, we set all the elements emitted by write head are equal to 0 to omit writing operation, and adaptively learn the read head in reading operation as the final weights. Different from the approaches in [33], [23], [1], [6], we don't implicitly obtain weights by dot-producting the feature vectors with a key vector. Conversely, we explicitly learn the weights from the whole feature vectors in the feature sets. As a consequence, our weighting is global-content-based and set-aware. Let $\{f_i\}$ be a face feature set and $\{s_i\}$ be the corresponding significance, which will be learnt adaptively. A softmax operation is applied on s_i to form normalized weights ω_i. The operation and the reformulated $E_\omega\left[f(x)\right]$ in Eq.(6) are as follows.

$$\omega_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)} \tag{7}$$

$$E_\omega\left[f(x)\right] = \sum_i [\frac{\exp(s_i)}{\sum_j \exp(s_j)} \cdot f_i] \tag{8}$$

Obviously, the weights $\{\omega_i\}$ and the $\{s_i\}$ are differentiable, allowing them to be learnt end-to-end.

C. Network Architecture

We introduce the lightened CNN [32] as our primary network. It contains 29 convolution layers with residual blocks and Max-Feature-Map (MFM) operations. Based on the primary network, ASML and Memory Attention Weighting are coupled for video face recognition. Moreover, we find Softmax is an important supervised signal for our method. Thus we obtain the following object function:

$$L = \lambda_1 Softmax + \lambda_2 \mathrm{ASML} \tag{9}$$

where λ_1 and λ_2 are trade-offs between these two terms. The softmax function is used for standard face recognition tasks, and ASML penalizes to increase the correlations and the discrepancies between the probability distributions of similar and dissimilar face sets respectively.

III. Experiments

A. Datasets

YouTubeFace (YTF) [31] contains 3,425 unconstrained videos varying from 48 to 6070 frames of 1,595 different subjects download from YouTube. This dataset is constructed for video face verification task with the standard testing protocol to perform ten-fold cross validation tests.

YouTube Celebrities (YTC) [11] is composed of 1,910 video clips of 47 subjects, which is also collected from YouTube website. It is a quite challenging dataset for video face identification task due to the high compression rate and large appearance variations. Following the standard evaluation protocol, five-fold cross validation is conducted on it.

Celebrity-1000 (C-1000) [14] is a large-scale unrestricted video dataset downloaded from YouTube and Youku. It is designed for video face identification task and consisted of 159,726 video sequences of 1000 subjects. There are two types of protocols on this dataset: close-set and open-set. In the close-set protocol, four overlapping subsets are created with an incremental number of subjects: 100, 200, 500 and 1000. In the open-set protocol, 200 subjects are used for training, the rest 800 subjects are used as gallery and probe with four settings: 100, 200, 400 and 800 subjects.

Table I
COMPARISON OF VERIFICATION ACCURACY RATE (±STANDARD VARIATION) (VAR±SD) WITH OTHER STATE-OF-THE-ART METHODS ON THE YOUTUBEFACE DATASET.

Method	VAR±SD (%)
DeepFace[27]	91.40±1.10
DeepID2+[24]	93.20±0.20
Sparse ConvNet[25]	93.50
VGG-CRF-SME[22]	93.80±1.30
Wen et al.[30]	94.90
TBE-CNN[5]	94.96±0.31
FaceNet[21]	95.12±0.39
NAN[33]	95.72±0.64
VGG-Face[19]	97.30
Our(Baseline)	95.54±1.12
Our(Softmax)	96.74±0.78
Our(Softmax+ASML)	**97.58±0.79**

Table II
AVERAGE RANK-1 ACCURACY (%) ON THE YOUTUBE CELEBRITIES DATASET WITH STANDARD 5-FOLD CROSS VALIDATION.

Method	Rank-1 (%)
LMKML[17]	78.20
MMDML[16]	78.50
MSSRC[18]	80.75
SFSR[35]	85.74
RRNN[13]	86.60
CRG[3]	86.70
VGG-Face[19]	93.62
Our(Baseline)	94.18
Our(Softmax)	95.39
Our(Softmax+ASML)	**97.52**

B. Training Schema

The network training includes three stages. First, the primary CNN is trained on the MS-Celeb-1M dataset followed by the instructions in [32]. Then we remove the last fully connected layer and append a dataset-specific fully connected layer with randomly initialized. Fixing all the other network parameters and training parameters, we fine-tune this layer with 1e-3 learning rate under the softmax supervised signal. At the third stage, the whole network framework is fine-tuned on the above three datasets. In this stage, for all experiments, we set the learning rate to 1e-4 with fixed policy and initialize the memory attention weights to be equal with the summation to be 1. The coefficients of momentum and weight decay are set to 0.9 and 5e-4 respectively. The trade-off parameters λ_1 (for Softmax term) and λ_2 (for ASML term) are assigned to 1 and 0.01, the margin α is set to 25. Note that the face images in YTC are poor quality and low-resolution (20×20). Accordingly, they may lead to over-fitting if the large network parameters are leant. To address this problem, we enlarge the weight decay to 5e-3 and the margin to 60 while shrink λ_2 to 0.004 for experiments on YTC. All the colored face images are transformed to gray-scale and normalized to 144×144. Then they are randomly cropped into 128×128 and mirrored with 50% probability to augment the training data. The batch size is set to 72, and each batch contains three equal-sized subsets of face images to form a triple-set, including an anchor subset X, a positive subset Y and a negative subset Z. We use the hard sampling approach to construct these triple-sets.

C. Results on YouTubeFace

We evaluate the video face verification performance of our method on the YTF dataset. Following the standard verification protocol (described in Sec.III-A), we report the average Verification Accuracy Rate (VAR) under 10-fold cross validation.

As shown in Table I, we compare our method with DeepFace [27], VGG-Face [19], DeepID2+ [24], Sparse ConvNet [25], VGG-CRF-SME [22], Wen et al. [30], TBE-CNN [5], FaceNet [21], GoogleNet [26] and NAN [33]. Our proposed method obtains 97.58% VAR, which beats the recently popular approaches with large margins, such as DeepFace, DeepID2+ and FaceNet. It outperforms the previous best method VGG-Face (97.3%) and the feature aggregation method NAN (95.72%). It worth nothing that VGG-Face extracts each face feature by averaging 30 cropped patches (three scales, five random positions, and horizontal mirror), whereas our method only uses 1 patch.

D. Results on YouTube Celebrities

In this subsection, we report the results of our proposed method on YTC, following the standard 5-fold cross validation for identification task.

Table II reports the comparisons of our method with Localized Multi-Kernel Metric Learning (LMKML) [17], Multi-Manifold Deep Metric Learning (MMDML) [16], Mean Sequence Sparse Representation-based Classification (MSSRC) [18], Simultaneous Feature and Sample Reduction (SFSR) [35], Recurrent Regression Neural Network (RRNN) [13], Covariate-Relation Graph (CRG) [3] and VGG-Face [19]. It is obvious that Softmax+ASML obtains the best performance compared with other methods. What is more, It improves the accuracy of non CNN-based methods by more than 11% and reduces the error rate of CNN-based methods by about 60%. This significant improvement indicates Softmax+ASML can drive CNNs to generate more robust and discriminative features for VFR.

E. Results on Celebrity-1000

We compare the performance of our method with other state-of-the-art methods, including Multi-task Joint Sparse Representation (MTJSR) [34], Eigen Probabilistic Elastic Part (Eigen-PEP) [12], Deep Extreme Learning Machines (DELM) [28], GoogleNet [26] and Neural Aggregation Network (NAN) [33] on the close-set and open-set protocols of C-1000.

close-set protocol: The evaluation results on the close-set setting are shown in the left of Table III. The proposed Softmax+ASML outperforms the state-of-the-art methods on Subject-100 (90.84% vs 90.44%), Subject-200 (86.71% vs 83.33%), Subject-500 (84.09% vs 82.27%) and Subject-1000 (81.92% vs 77.17%). It improves the performance of the baseline network by large margins, because ASML exploits the typical information in videos based on memory attention mechanism.

Table III
COMPARISON OF RANK-1 ACCURACY (%) WITH OTHER STATE-OF-THE-ART METHODS AND DIFFERENT SUPERVISED SIGNALS ON THE CELEBRITY-1000 DATASET.

Method	Close-Set				Open-Set			
	100	200	500	1000	100	200	400	800
MTJSR[34]	50.60	40.80	35.46	30.04	46.12	39.84	37.51	33.50
DELM[28]	49.80	45.21	38.88	28.83	-	-	-	-
Eigen-PEP[12]	50.60	45.02	39.97	31.94	51.55	46.15	42.33	35.90
NAN[33]	90.44	83.33	82.27	77.17	88.76	85.21	82.74	79.87
Our(Baseline)	87.25	81.70	78.37	74.76	86.77	82.38	81.12	76.87
Our(Softmax)	88.45	84.01	80.80	79.12	87.94	83.17	82.28	78.00
Our(Softmax+MMD)	89.24	84.97	82.99	80.48	88.33	84.15	82.87	78.64
Our(Softmax+RMD)	89.64	85.36	83.07	80.87	88.72	84.55	83.25	78.94
Our(Softmax+MSML)	90.04	86.13	83.78	81.37	89.11	85.35	83.54	79.43
Our(Softmax+ASML)	**90.84**	**86.71**	**84.09**	**81.92**	**89.88**	**85.94**	**83.83**	**80.02**

open-set protocol: The right side of Table III displays the results on the open-set protocol. DELM is not evaluated on this protocol in the published paper [28], thus we remove it from this table. As expected, Softmax+ASML achieves the best results, performing 89.88%, 85.94%, 83.83% and 80.02% for the settings of 100, 200, 400 and 800 subjects respectively. It indicates that our method has superior generalization capability and can better recognize the subjects that are not seen in training.

F. Ablation Study

In this subsection, we analysis the effect of different supervised signals (see Sec.II-A) on Celebrity-1000 dataset. The results are presented in the lower part of Table III. As shown in this Table, RMD has better performance than MMD, and ASML is superior to MSML on both the close-set and open-set protocols. It indicates that the memory attention weighting for sample biases correction is effective for video face recognition. Then, comparing with ASML and RMD, the former has an improvement of Rank-1 accuracy by average 0.96%, which is consistent with the results between MSML and MMD. It means that pulling the feature distributions of different subjects (MSML and ASML) is effective for video face recognition.

IV. CONCLUSION

In this paper, we have introduced an Attention-Set based Metric Learning (ASML) method for video face recognition. By exploiting memory attention weighting, our proposed method corrects the sample biases in the training face videos, and can be embedded into CNNs seamlessly with end-to-end training. Specifically, ASML drives CNNs to generate more discriminative face representations containing small intra-set and large inter-set distance. Extensive experiments on three popular benchmarks have demonstrated that our proposed method is superior to the other state-of-the-art approaches.

ACKNOWLEDGMENT

This work is partially funded by National Natural Science Foundation of China (Grant No. 61622310, 61473289), the State Key Development Program (Grant No. 2016YFB1001001) and Beijing Municipal Science and Technology Commission (No.Z161100000216144).

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *In ICLR*, 2015.

[2] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *ECCV*, 2012.

[3] Z. Chen, B. Jiang, J. Tang, and B. Luo. Image set representation and classification with covariate-relation graph. In *ACPR*, 2015.

[4] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *ICALT*, 2008.

[5] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *arXiv preprint arXiv:1607.05427*, 2016.

[6] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[7] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, 2011.

[8] M. Hayat, M. Bennamoun, and S. An. Learning non-linear reconstruction models for image set classification. In *CVPR*, 2014.

[9] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *CVPR*, 2015.

[10] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, 2015.

[11] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008.

[12] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. In *ACCV*, 2014.

[13] Y. Li, W. Zheng, and Z. Cui. Recurrent regression for face recognition. *arXiv preprint arXiv:1607.06999*, 2016.

[14] L. Liu, L. Zhang, H. Liu, and S. Yan. Toward large-population face identification in unconstrained videos. *IEEE TCSVT*, 24(11):1874–1884, 2014.

[15] J. Lu, G. Wang, W. Deng, and P. Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*, 2014.

[16] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, 2015.

[17] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, 2013.

[18] E. G. Ortiz, A. Wright, and M. Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *CVPR*, 2013.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[20] B. Schlkopf, J. Platt, and T. Hofmann. A kernel method for the two-sample-problem. *In NIPS*, 2008.

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

[22] M. Shao, Y. Zhang, and Y. Fu. Collaborative random faces-guided encoders for pose-invariant face representation learning. *IEEE T NEUR NET LEAR*, 2017.

[23] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *NIPS*, 2015.

[24] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015.

[25] Y. Sun, X. Wang, and X. Tang. Sparsifying neural network connections for face recognition. In *CVPR*, 2016.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[28] M. Uzair, F. Shafait, B. Ghanem, and A. Mian. Representation learning with deep extreme learning machines for efficient image set classification. *NEURAL COMPUT APPL*, pages 1–13, 2015.

[29] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen. Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. In *CVPR*, 2015.

[30] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[31] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

[32] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2016.

[33] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. *arXiv preprint arXiv:1603.05474*, 2016.

[34] X.-T. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *IEEE TIP*, 21(10):4349–4360, 2012.

[35] M. Zhang, R. He, D. Cao, Z. Sun, and T. Tan. Simultaneous feature and sample reduction for image-set classification. In *AAAI*, 2016.