2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Regularized Metric Adaptation for Unconstrained Face Verification

Boyu Lu, Jun-Cheng Chen, and Rama Chellappa
Center for Automation Research, UMIACS
University of Maryland
College Park, Maryland, 20740
Email: bylu@umiacs.umd.edu, pullpull@cs.umd.edu,rama@umiacs.umd.edu

***Abstract*—In this work, we propose a metric adaptation method for set-based face verification and evaluate it on the newly released IARPA Janus Benchmark A (IJB-A) dataset and its extended version, the Janus Challenging Set 2 (CS2). A template-specific metric is trained to adaptively learn the discriminative information in test templates and the negative training set, which contains subjects that are mutually exclusive to subjects in test templates. The proposed regularized joint Bayesian metric learning framework not only alleviates the over-fitting problem but also provides a way to efficiently reduce the model size. We also analyze the selection of the compact and representative negative set to speed up the training time and to reduce storage space. Experiments on the IJB-A and CS2 datasets yield promising results.**

I. Introduction

Face verification research has been one of the active research areas in computer vision community for decades. Although performance on the well-known unconstrained face dataset, the Labeled Faces in the Wild (LFW) [10][13], has been pushed to surpass human performance by deep convolutional neural network approaches [18][16], most faces in the LFW dataset are in frontal or near-frontal poses. Therefore, the problem of unconstrained face verification under extreme pose, illumination, and expression variations is still unsolved. Recently, the IARPA Janus Benchmark A (IJB-A) dataset [12] was released to address these problems which contains more challenging unconstrained acquisition conditions, including full pose and illumination variations, aging and occlusion. In addition, the protocol for IJB-A verification is set/template-based, and each set/template contains a mixture of images or frames from multiple videos of the same person. Different from the traditional image-to-image or single video-to-video comparisons, this new protocol is more challenging and practical because faces from heterogeneous sources introduce larger variations within the templates.

In this paper, we propose a metric adaptation method for the set-based face verification problem. Given a pair of templates, the idea of metric adaptation is to learn a template-specific metric by utilizing the intra-information between features in one template and the inter-information between the template and the negative set (*i.e.*, the negative set is a set of samples from subjects who are mutually exclusive to the test data.). In principle, this is similar to the one-shot approach [21] used for one-to-one verification where intra-information cannot be exploited.

In general, the proposed regularized joint Bayesian metric learning framework alleviates the over-fitting problem. In addition, it provides a way to significantly reduce the model size without much degradation in performance. We also analyze the selection of the negative set to reduce its size and to accelerate the metric learning process. Extensive experiments on IJB-A and CS2 datasets yield promising results compared to other competitive methods.

The rest of the paper is organized as follows. A brief review of related works is presented in Section II. In Section III, we present the details of the proposed method and the strategy for selecting the negative set. We present the experimental results in Section IV and finally conclude in Section V.

II. Related Work

We briefly review several related works on metric learning for face verification problems.

The discriminative similarity measure is a key component in improving the performance for a face verification system. Taigman *et al.* [19] learned the Mahalanobis distance for face verification using the Information Theoretic Metric Learning (ITML) proposed in [6]. Weinberger *et al.* [20] proposed the Large Margin Nearest Neighbor (LMNN) metric which enforces the large margin constraint on the relative distances among all triplets of labeled training data (*i.e.* a triplet consists of an anchor sample, a positive sample with the same label as the anchor and a negative sample with a different label.). Chen *et al.* [2] proposed a joint Bayesian approach which models the joint distribution of a pair of face images directly instead of their difference vector, and the ratio of between-class and within-class probabilities is adopted as the similarity measure. Hu *et al.* [9] proposed a discriminative deep metric from hand-crafted features for face verification using a deep neural network. Huang *et al.* [11] proposed a projection metric which preserves the underlying manifold structure of the labeled training images. Schroff *et al.* [16] and Parkhi *et al.* [14] optimized the DCNN parameters based on triplet distance embedding loss which directly embeds the DCNN features into a discriminative subspace and presented promising results for face verification. Recently, Sankaranarayanan *et al.* [15] used triplet metric learning based on similarity instead of

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

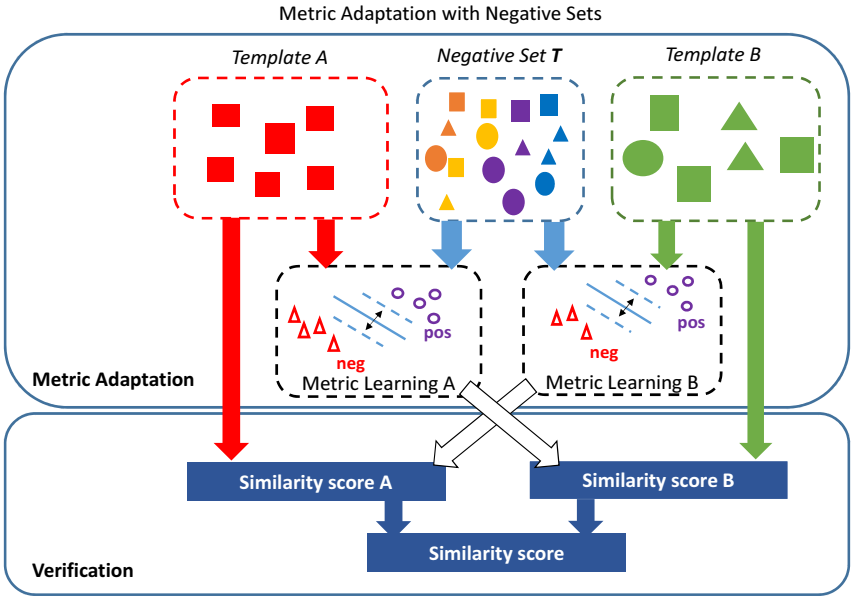


Fig. 1. The system overview of the proposed regularized metric adaptation method for unconstrained face verification.

a distance measure to achieve better results for the face verification task.

On the other hand, besides learning a metric from the training data, Wolf *et al.* [21] proposed the one-shot similarity (OSS) kernel based on a set of pre-selected reference images that are mutually exclusive to the pair of images being compared and training a discriminative classifier between the test images and the reference set. Guo *et al.* [7] followed the same paradigm and developed the one-shot similarity approach based on partial least square regressors to leverage the rich information of the high-dimensional feature obtained by concatenating Gabor [22], LBP [1], and HOG [5] features. In [8], Guo *et al.* extended their one-shot similarity approach using a sparse representation that does not require any training of classifiers between the reference set and test samples. Crosswhite *et al.* [4] developed a one-shot similarity framework based on linear support vector machines and deep convolutional features of faces and achieved competitive results for the unconstrained face verification task. Our approach falls in this category: we adapt the joint Bayesian metric to the one-shot similarity framework and achieve good results. Furthermore, we also demonstrate in Section IV that our regularized formulation can be used to reduce the model size but still yield comparable results to those obtained using the original metric.

III. Proposed Method

A. Regularized Joint Bayesian Metric Learning

The joint Bayesian metric learning has been shown to be effective for face verification [2], [3]. Its formulation can also be interpreted as the combination of two components: Mahalanobis distance and projected cosine similarity. In general, directly minimizing the hinge loss objective function usually results in a large model complexity and over-fitting problems due to a large number of parameters introduced by metric matrices. On the other hand, Euclidean distance and cosine similarity provide a good baseline performance on deep convolutional features [3] for the face verification task. In addition, Euclidean distance and cosine similarity have better generalization capability because they are not trained on a particular training set. The model size for Euclidean and cosine metric is also small since only the diagonal terms are non-zeros. Therefore, we add the regularization terms to enforce the learned metric matrices to stay close to identity matrices, since when both metric matrices are identity, the computation of the similarity scores reduces to the summation of the Euclidean distance and the cosine similarity.

Given a set of features $\mathbf{X}$, we construct positive pairs if both features belong to the same person and negative pairs otherwise. The goal of the metric learning is to increase the similarity score of positive pairs while decreasing the negative ones. We solve an optimization problem as follows:

$$\underset{\mathbf{W},\mathbf{V},b}{\operatorname{argmin}} \sum_{ij} \max\{0, \alpha - l_{ij}(b - d_{\mathbf{W}}(x_i, x_j) + 2s_{\mathbf{V}}(x_i, x_j))\} + \lambda_1 \|\mathbf{W} - \mathbf{I}\|_F^2 + \lambda_2 \|\mathbf{V} - \mathbf{I}\|_F^2 \quad (1)$$

where $d_{\mathbf{W}}(x_i, x_j) = (x_i - x_j)^T \mathbf{W}^T \mathbf{W}(x_i - x_j)$ is the Mahalanobis distance and $s_{\mathbf{V}}(x_i, x_j) = x_i^T \mathbf{V}^T \mathbf{V} x_j$ is the projected similarity. Both $\mathbf{W} \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \in \mathbb{R}^{d \times d}$ are the projection matrices. $l_{ij} = 1$ if $\{x_i, x_j\}$ is a positive pair and $l_{ij} = -1$, otherwise. b is the bias and α is the margin parameter. λ_1, λ_2 are the regularization parameters to control the regularization terms.

To solve the optimization problem in (1), we apply the

stochastic gradient descent (SGD) method as follows:

$$\begin{aligned}
\mathbf{W}_{t+1} &= \begin{cases} \mathbf{W}_t, & \text{if } l_{ij}\rho_{ij} \geq \alpha \\ \mathbf{W}_t - \tau(l_{ij}\mathbf{W}_t\mathbf{\Psi}_{ij} + \lambda_1(\mathbf{W}-\mathbf{I})), & \text{otherwise,} \end{cases} \\
\mathbf{V}_{t+1} &= \begin{cases} \mathbf{V}_t, & \text{if } l_{ij}\rho_{ij} \geq \alpha \\ \mathbf{V}_t + \tau(l_{ij}\mathbf{V}_t\mathbf{\Gamma}_{ij} + \lambda_2(\mathbf{V}-\mathbf{I})), & \text{otherwise,} \end{cases} \\
b_{t+1} &= \begin{cases} b_t, & \text{if } l_{ij}\rho_{ij} \geq \alpha \\ b_t + \tau l_{ij}, & \text{otherwise,} \end{cases}
\end{aligned} \tag{2}$$

where τ is the learning rate, $\mathbf{\Psi}_{ij} = (x_i - x_j)(x_i - x_j)^T$, $\mathbf{\Gamma}_{ij} = x_i x_j^T + x_j x_i^T$, $\rho_{ij} = b - d_{\mathbf{W}}(x_i, x_j) + 2s_{\mathbf{V}}(x_i, x_j)$. Note that the regularization term is updated only when the condition is violated instead of being updated for every iteration. In practice, this strategy significantly reduces the computation complexity but yields similar results.

B. Metric Adaptation with Negative Set

Given a negative training set T which has no overlapping subjects with the test set and a pair of test templates G and P, we adaptively learn two metric metrics for templates G and P as described below. The positive pairs are generated by every two features in G (*i.e.*, if the template only contains a single face image, we use the features extracted from the image and its horizontally flipped one.). On the other hand, the negative pairs are generated for every two features between G and T (*i.e.*, one in G, and the other one in T.). With a bunch of positive and negative pairs, we train the regularized metric for G by solving (1). Once the metric matrices are learned, we compute the similarity score $\rho_G(P, G) = b_G - d_{\mathbf{W}_G}(x_G, x_P) + 2s_{\mathbf{V}_G}(x_G, x_P)$, where x_G and x_P are the average of unit-normalized features for the template (*i.e.* the average used here is media sensitive: the features from the same video will be averaged first and then averaged with others.). Similarly, we train a metric for the template P and compute $\rho_P(P, G)$. Finally, the similarity score between G and P is computed as the weighted sum of the two scores: $s(P, G) = \beta\rho_G + (1 - \beta)\rho_P$ where β is the weight used to balance the two similarity scores and is determined as the ratio of the number of positive pairs in each template. The overview of the proposed method is illustrated in Figure 1.

C. Negative Set Selection

In general, a large negative set is preferred for metric adaptation since more diverse negative pairs help to learn a better metric. However, since metric adaptation is conducted during testing time, it is essential to reduce the size of the negative set to speed up the computation. One simple solution is to directly average and normalize the features by subjects and use the averaged features as the negative set. However, since the training set contains some faces which may be badly aligned or in extreme pose or illumination conditions, directly averaging them with other good features introduces errors and degrades the performance. We develop a strategy to identify outliers based on the results of K-means clustering and only use the good features for averaging. First, the mean feature of each subject is used to initialize the K-means algorithm, K is set as the number of subjects in the negative set, and then we apply the K-means algorithm on the entire negative set. In the best situation, all the features should be assigned to the cluster corresponding to their ground truth labels. If some features are assigned to the clusters of other subjects, these features are potential outliers to their own subjects. Nevertheless, if the subjects contain very few features, it is possible that all the features in the subjects are assigned to other subjects. In this case, we should preserve all the features in the subjects. The detailed steps are summarized in Algorithm 1.

Algorithm 1 Negative Set Selection

Input: Original Negative Set X, class labels for all the features in X.

Output: Representative negative set X_r.

1. Mean selection: For each subject i, compute the mean point x_{M_i}

2. Representative feature selection: Apply the K-means algorithm on the entire set X, using all the x_{M_i} obtained from step 1 for initialization. For each feature, compare its new cluster index with its true label. Preserve the consistent ones.

3. Outliers removing: Remove the non-consistent features. If there is no consistent feature for certain subjects, preserve all the features.

4. Representative features averaging: Average the remaining features in each subject to get the final negative set X_r.

IV. Experimental Results

In this section, we evaluate the proposed approach on the challenging IARPA Janus Benchmark A (IJB-A) and its extended version, the Janus Challenging Set 2 (CS2). Some alternative methods are compared and the receiver operating characteristic curves (ROC) are used to measure the performance for different algorithms. We also discuss the reduction of model size and the selection of the negative set.

A. Experiment Setup

The DCNN features used in all the experiments of this work are the pool5 features extracted by the deep convolutional network proposed in [3] which consists of ten convolutional layers, five pooling layers and one fully connected layer and is trained using the CASIA-WebFace dataset [23]. The dimensionality of the pool5 features is 320. Media averaging pooling followed by unit-normalization for the feature vectors are used as the preprocessing steps after feature extraction [4].

For the parameters used in (1), we set margin $\alpha = 0.001$, regularization parameters $\lambda_1 = \lambda_2 = 0.01$, and the learning rate $\tau = 0.01$. In general, a large margin results in a more strict condition for $l_{ij}\rho_{ij} \geq \alpha$ in (2), where the condition is easier to be violated and the metric will be updated very often. This may discourage the metric from learning the hard positives or negatives. Therefore, we set the margin to a relatively small number so that the metric is updated based

Method	Negative Set Usage, Size	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
Cosine	No	0.598±0.078	0.802±0.055	0.945±0.009
JBML	Yes, during training period, about 10,000	0.655±0.072	0.836±0.028	0.955±0.006
SVM-TA-v0 [4]	Yes, during adaptation period, N/A	N/A	0.939±0.013	N/A
SVM-TA-v1 [4]	Yes, during adaptation period, 332	0.723±0.034	0.874±0.012	0.956±0.006
SVM-TA-v1 [4]	Yes, during adaptation period, about 10,000	0.757±0.048	0.888±0.013	0.956±0.007
RMA	Yes, during adaptation period, 332	0.763±0.037	0.887±0.014	0.959±0.005

TABLE I
VERIFICATION RESULTS ON IJB-A DATASET. THE RESULTS ARE AVERAGED OVER 10 SPLITS. THE RESULTS OF SVM-TA-V0 IN THE THIRD ROW ARE DIRECTLY CITED FROM THE ORIGINAL PAPER. THE RESULTS OF SVM-TA-V1 ARE IMPLEMENTED BY US.

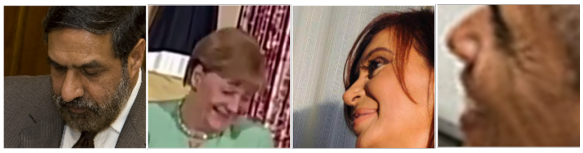


Fig. 3. Sample images in IJB-A dataset.

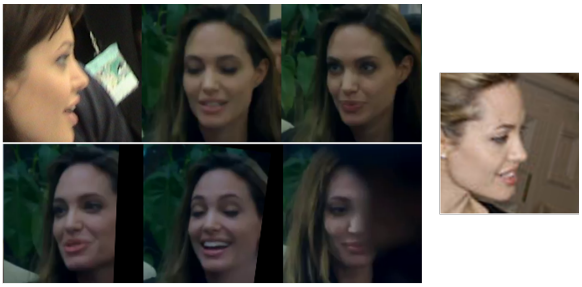


Fig. 4. Sample pair that is correctly classified by RMA while mis-classified by JBML.

on the hard negative/positive pairs. This idea is similar to the hard negative/positive mining strategy which is widely used in metric learning and has proven to be effective [17], [16], [15]. The learning rate and the regularization parameter are determined based on cross validation. We initialize $\mathbf{W_0} = \mathbf{V_0} = \mathbf{I}$ and b_0 is learned using only the negative set during the training period. The size of negative set is 332 which is the number of subjects in the set. In our experiments, all the possible positive and negative pairs are used to learn the metric for five epochs because the size of the negative set and the testing templates are small. The weight used to balance the two similarity scores is set as the ratio of the number of positive pairs in each template.

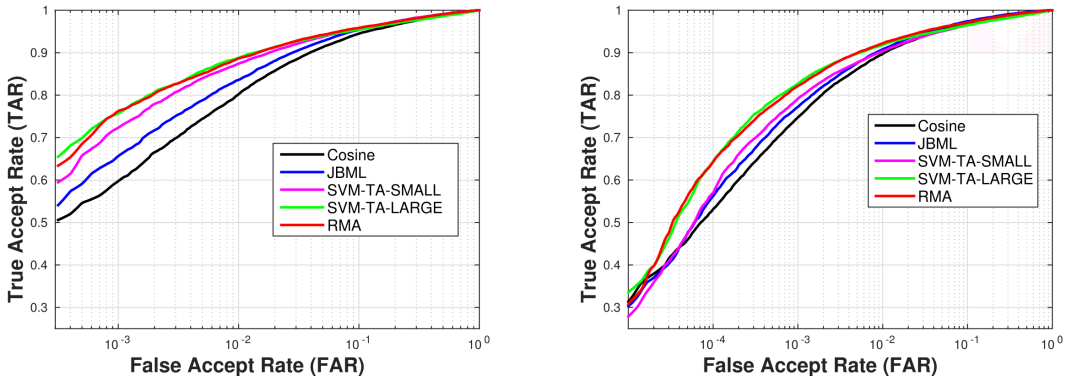
B. Evaluation on IJB-A and CS2 Datasets

Both IJB-A and JANUS CS2 datasets contain 500 subjects with 5,397 images and 2,042 videos. The datasets are divided into training sets which contain 333 subjects, and test sets which contain 167 subjects. Based on the different training/ test set division, ten splits are generated. Some sample images are shown in Figure 3. The training sets are shared for both datasets. For the test set, JANUS CS2 contains about 167 gallery templates and 1763 probe templates. All pairs of gallery-to-probe templates are used for verification. The IJB-A evaluation protocol selects around 11,748 hard pairs of gallery-to-probe templates (1,756 positive and 9,992 negative pairs) from JANUS CS2.

We compare the results of the proposed regularized metric adaptation (RMA) approach with two baseline methods, the cosine similarity without metric learning and the joint Bayesian metric learning (JBML) without metric adaptation. The cosine similarity method is unsupervised and does not require any training set while JBML is trained using the training data of IJB-A and JANUS CS2 during the training period and the trained model is then applied in the testing phase. We also compare our results with the recently proposed SVM-based template adaptation (SVM-TA) method [4], which requires a large negative set in testing phase for template adaptation. We cite the results from [4] as SVM-TA-v0. We also follow the same preprocessing steps and use the same parameters described in [4] for our implemented features as SVM-TA-v1 for comparison. The main difference comes from the DCNN features used in both works where in [4] the network is trained using the VGG face dataset which contains more face images (around 2.6 million faces) than the CASIA-WebFace dataset (around 500K faces) used by us.

Figure 2(a) shows the ROC curves for the IJB-A dataset. Table IV shows the True Acceptance Rate (TAR) when False Alarm Rate (FAR) at $10^{-3}, 10^{-2}, 10^{-1}$. The results are averaged over 10 splits. It is observed that the proposed method RMA shows better results than other non-adaptation baselines especially in the low FAR region. Figure 4 shows an example pair that is correctly classified by RMA, yet mis-classified by JBML at $FAR = 10^{-2}$. It demonstrates the effectiveness of the metric adaptation approach for the hard case, where extreme poses and occlusions are present. Notice that two versions of the SVM-TA-v1 results are reported based on whether a small or a large negative set is used. We outperform SVM-TA-v1 when using the same negative set while perform comparably when SVM-TA-v1 uses a larger negative set. It demonstrates that metric learning can fully exploit the discriminative information in a relatively small negative set.

Figure 2(b) shows the ROC curves for the CS2 dataset. Table IV-B shows the performance of different methods on the CS2 dataset. Results are averaged over 10 splits. As an



(a) (b)

Fig. 2. ROC curves for IJB-A and CS2 dataset. The results are averaged over 10 splits. SVM-TA-SMALL means using a small negative set and SVM-TA-LARGE means using a large negative set where SVM-TA refers to our implementation, SVM-TA-v1.

Method	Negative Set Usage, Size	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
Cosine	No	0.748±0.031	0.898 ±0.010	0.945 ±0.003
JBML	Yes, during training period, about 10,000	0.773±0.040	0.908±0.007	0.974±0.004
SVM-TA-v1 [4]	Yes, during adaptation period, 332	0.792±0.018	0.904±0.007	0.965±0.004
SVM-TA-v1 [4]	Yes, during adaptation period, about 10,000	0.827±0.014	0.918± 0.007	0.965±0.003
RMA	Yes, during adaptation period, 332	0.822±0.019	0.922±0.008	0.971±0.002

TABLE II
VERIFICATION RESULTS ON CS2 DATASET. THE RESULTS ARE AVERAGED OVER 10 SPLITS.

extended version of IJB-A dataset, the CS2 dataset compares all the possible pairs in the gallery and probe sets. The baseline for CS2 is higher than for the IJB-A dataset which makes it more difficult to improve from the baseline. The proposed RMA still outperforms the non-adaptation method by 2% at $FAR = 10^{-2}$ and 5% at $FAR = 10^{-3}$. SVM-TA-v1 with the large negative set still yields comparable results. However, when using the same negative set, it can hardly improve the performance from the non-adaptation baselines.

C. Model Size Reduction

When the model learned by the metric adaptation needs to be saved for future use (*e.g.*, the subject is enrolled in the database.), it is useful to reduce the model size as small as possible for practical use. The original model requires $\mathcal{O}(n^2)$ storage space where n is the dimension of the data sample. Since the model is template-specific, the whole model size for a dataset will be proportional to the number of unique templates which is usually very large. We reduce the original model size to $\mathcal{O}(n)$ by taking only the diagonal of $\mathbf{W}$ and the transformed feature $\mathbf{V}^T\mathbf{V}x$ for each template. The similarity is then computed as $\rho_G(x_G, x_P) = b_G - (x_G - x_P)^T diag(\mathbf{W}_G)^2(x_G - x_P) + 2x_P\mathbf{V}_G^T\mathbf{V}_G x_G$ and similarly for $\rho_P(x_G, x_P)$. The reason why we keep the diagonal elements of $\mathbf{W}$ is that as we enforce a regularization term in (1), which guarantees that the elements on the diagonal preserve the most information as compared to other off-diagonal elements. The results with and without model size reduction are listed in Table III. From the table, the performance only decreases by a small margin while the whole model size is significantly reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$.

D. Negative Set Selection Analysis

The size of the negative set significantly influences the adaptation time as well as the storage space. It is desired to keep a relatively small negative set while maintaining a similar performance as the large one. We investigate and compare different strategies to reduce the size, including (1) ***Random*** where a media feature (*i.e.* features from the same media are averaged) for each subject is randomly selected into the negative set, (2) ***Naive K-means*** where the media average feature for each subject (*i.e.* features from the same media are averaged first and then different media from one subject are averaged) is used as the negative set, (3) ***Naive K-medoids*** where the 1-medoid of all the media features of each subject is taken into the negative set, (4) ***Outlier Removed K-means*** means the method described in Algorithm 1, and (5) ***Outlier Removed K-medoids*** means the similar strategy described as *Outlier Removed K-means* but K-means is replaced by K-medoids.

Table IV summarizes the results of different methods using RMA on IJB-A verification split 1. It shows that methods based on K-means outperform K-medoids based method and randomly selection by a large margin. It shows that by averaging different media in one subject, we obtain more discriminative information than just including a single media feature. The *Outlier Removed Kmeans* performs slightly better than *Naive Kmeans* at $FAR = 10^{-2}$.

Model Size	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
$\mathcal{O}(n)$	0.746±0.041	0.878±0.016	0.956±0.005
$\mathcal{O}(n^2)$	0.763±0.037	0.887±0.014	0.959±0.005

TABLE III
THE RESULTS FOR THE MODEL SIZE REDUCTION WHICH ARE AVERAGED OVER 10 SPLITS.

Method	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
Random	0.683	0.848	0.943
Naive K-means	0.773	0.886	0.952
Outlier Removed K-means	0.770	0.890	0.953
Naive K-medoids	0.672	0.851	0.946
Outlier Removed K-medoids	0.673	0.851	0.947

TABLE IV
NEGATIVE SET SELECTION. IT SHOWS THE RESULTS OF DIFFERENT STRATEGIES FOR THE SPLIT 1 OF THE IJB-A FACE VERFICATION.

V. CONCLUSION

In this paper, we proposed a regularized metric adaptation approach to learn a template-specific metric for the set-based face verification problem.

Extensive experiments on the newly released IARPA Janus Benchmark A (IJB-A) and CS2 dataset demonstrate the effectiveness of the proposed method for unconstrained face verification when the negative set is used. In addition, the proposed approach can be used to significantly reduce the model size while still yielding comparable performance to the original model. Analysis shows the importance of the negative set selection on the verification performance. A K-means based method can efficiently construct a compact and representative negative set.

ACKNOWLEDGMENTS

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer, 2012.

[3] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. *arXiv preprint arXiv:1508.01722*, 2015.

[4] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman. Template Adaptation for Face Verification and Identification. *ArXiv e-prints*, Mar. 2016.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine learning*, pages 209–216, 2007.

[7] H. Guo, W. R. Schwartz, and L. S. Davis. Face verification using large feature sets and one shot similarity. In *International Joint Conference on Biometrics*, pages 1–8. IEEE, 2011.

[8] H. Guo, R. Wang, J. Choi, and L. S. Davis. Face verification using sparse representations. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–44. IEEE, 2012.

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

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report.

[11] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on Grassmann manifold with application to video based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 140–149, 2015.

[12] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939. IEEE, 2015.

[13] E. Learned-Miller, G. Huang, A. RoyChowdhury, H. Li, G. Hua, and G. B. Huang. Labeled faces in the wild: A survey. 2016.

[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *British Machine Vision Conference*, 2015.

[15] S. Sankaranarayanan, A. Alavi, and R. Chellappa. Triplet similarity embedding for face verification. *arXiv preprint arXiv:1602.03418*, 2016.

[16] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[17] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, and S. Z. Li. Constrained deep metric learning for person re-identification. *arXiv preprint arXiv:1511.07545*, 2015.

[18] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *arXiv preprint arXiv:1412.1265*, 2014.

[19] Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. In *British Machine Vision Conference*, pages 1–12, 2009.

[20] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2005.

[21] L. Wolf, T. Hassner, and Y. Taigman. Similarity scores based on background samples. In *Asian Conference on Computer Vision*, pages 88–97. 2010.

[22] S. Xie, S. G. Shan, X. L. Chen, and J. Chen. Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Transactions on Image Processing*, 19(5):1349–1361, 2010.

[23] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.