2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition

Template Adaptation for Face Verification and Identification

Nate Crosswhite[1], Jeffrey Byrne[1], Chris Stauffer[2],
Omkar Parkhi[3], Qiong Cao[3] and Andrew Zisserman[3]
[1] Systems and Technology Research, Woburn MA USA
[2] Visionary Systems and Research, Framingham, MA USA
[3] Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford UK

***Abstract*— Face recognition performance evaluation has traditionally focused on one-to-one verification, popularized by the Labeled Faces in the Wild dataset [1] for imagery and the YouTubeFaces dataset [2] for videos. In contrast, the newly released IJB-A face recognition dataset [3] unifies evaluation of one-to-many face identification with one-to-one face verification over *templates*, or sets of imagery and videos for a subject. In this paper, we study the problem of template adaptation, a form of transfer learning to the set of media in a template. Extensive performance evaluations on IJB-A show a surprising result, that perhaps the simplest method of template adaptation, combining deep convolutional network features with template specific linear SVMs, outperforms the state-of-the-art by a wide margin. We study the effects of template size, negative set construction and classifier fusion on performance, then compare template adaptation to convolutional networks with metric learning, 2D and 3D alignment. Our unexpected conclusion is that these other methods, when combined with template adaptation, all achieve nearly the same top performance on IJB-A for template-based face verification and identification.**

I. Introduction

Face recognition performance using deep learning has seen dramatic improvements in recent years. Convolutional networks trained with large datasets of millions of images of thousands of subjects have shown remarkable capability of learning facial representations that are invariant to age, pose, illumination and expression (A-PIE) [4], [5], [6], [7], [8], [9]. These representations have shown strong performance for recognition of imagery and video in-the-wild in unconstrained datasets, with recent approaches demonstrating capabilities that exceed human performance on the well known Labeled Faces in the Wild dataset [1].

The problem of face recognition may be described in terms of face verification and face identification. Face verification involves computing a one-to-one similarity between a probe image and a reference image, to determine if two image observations are of the same subject. In contrast, face identification involves computing a one-to-many similarity between a probe media and a gallery of known subjects in order to determine a probe identity. Face verification is important for access control or re-identification tasks, and face identification is important for watch-list surveillance or forensic search tasks.

Face recognition performance evaluations have traditionally focused on the problem of face verification. Over the past fifteen years, face datasets have steadily increased in size in terms of number of subjects and images, as well as complexity in terms of controlled vs. uncontrolled collection and amount of A-PIE variability [10]. The Labeled Faces in the Wild dataset [1] contains 13233 images of 1680 subjects, and compares specific pairs of images of subjects to characterize 1:1 verification performance. Similarly, the YouTubeFaces dataset [2] contains 3425 videos of 1595 subjects, and compares pairs of videos of subjects for verification. These datasets have set the established standard for face recognition research, with steadily increasing performance [11], [5], [6], [4]. However, the imagery in LFW was constructed with a well known near-frontal selection bias, which means evaluations are not predictive of performance for large in-the-wild pose variation. In fact, recent studies have shown that while algorithm performance for near frontal recognition is equal to or better than humans, performance of automated systems at the extremes of illumination and pose are still well behind human performance [12].

The IJB-A dataset [3] was created to provide the newest and most challenging dataset for both verification and identification. This dataset includes both imagery and video of subjects manually annotated with facial bounding boxes to avoid the near frontal bias, along with protocols for evaluation of both verification and identification. Furthermore, this dataset performs evaluations over *templates* [13] as the smallest unit of representation, instead of image-to-image or video-to-video. A template is a set of all media (images and/or videos) of a subject that are to be combined into a single representation suitable for matching. Template based representations are important for many face recognition tasks, which take advantage of an historical record of observations to further improve performance. For example, a template provides a useful abstraction to capture the mugshot history of a criminal for forensic search in law enforcement, or lifetime enrollment images for visa or driver's licenses in civil identity credentialing for improved access control. Biometric templates have been studied for face recognition, where performance on older algorithms have increased given

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under contract number 2014-14071600010. 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 ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.11

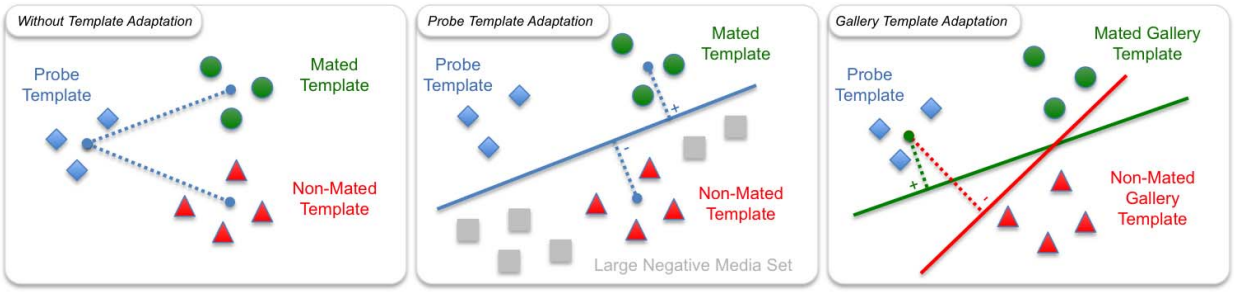


Fig. 1. Template Adaptation Overview. (left) Without template adaptation, the distance from the probe template to the mated subject ID and non-mated subject ID templates is about equal (dotted blue lines) resulting in incorrect equal similarity. (middle) With probe adaptation, a max-margin classifier (solid blue line) is trained to separate the probe template features from a large negative media feature set, which increases the mated template similarity (e.g. positive SVM margin from probe to green mated template shown with positive labeled dotted blue line) and decreases the non-mated (e.g. negative margin to red non-mated template shown with negative labeled dotted blue line). (right) With gallery adaptation, a max-margin classifier is trained for each gallery template independently (e.g. red solid line for non-mated gallery and green solid line for mated gallery templates, not shown are all other gallery templates affecting the optimal hyperplanes) to separate each gallery template features from all other gallery templates without considering a large negative set. The result is a correctly decreased similarity as a large negative margin between the probe and non-mated template (negative labeled dotted red line).

an historical set of images [13]. The IJB-A dataset is the only public dataset that enables a controlled evaluation of template-based verification and identification at the extremes of pose, illumination and expression.

In this paper, we study the problem of *template adaptation*. Template adaptation is an example of transfer learning, where the target domain is defined by the set of media of a subject in a template. In general, transfer learning includes a source domain for feature encoding of subjects trained offline, and a specific target domain with limited available observations of new subjects. In the case of template adaptation, the source domain may be a deep convolutional network trained offline to predict subject identity, and the target domain is the set of media in templates of never before seen subjects. In this paper, we study perhaps the simplest form of template adaptation based on deep convolutional networks and one-vs-rest linear SVMs. We combine deep CNN features trained offline to predict subject identity, with a simple linear SVM classifier trained at test time using all media in a template as positive features to classify each new subject.

Extensive evaluation of template adaptation on the IJB-A dataset has generated surprising results. First, template adaptation outperforms all top performing techniques in the literature: convolutional networks combined with triplet loss similarity [6], [4], [14], joint Bayesian metric learning [15], pose specialized networks [16], 2D alignment [4] and novel convolutional network architectures [17]. Second, template adaptation when combined with these other techniques results in nearly equivalent performance. Third, we show a clear tradeoff between the size of a template (e.g. the number of unique media in the template) and performance, which leads to the conclusion that if the average largest template size is big enough, then a simple template adaptation strategy is the best choice for both verification and identification.

II. Related Work

The top performing approaches for face verification on Labeled Faces in the Wild [1] and YouTubeFaces [2] are all based on convolutional networks. VGG-Face is the application of the VGG-16 convolutional network architecture [18] trained on a curated dataset of 2.6M images of 2622 subjects. This representation includes triplet loss embedding and 2D alignment for normalization to provide state of the art performance. FaceNet [6] applied the inception CNN architecture [19] to the problem of face verification. This approach included metric learning to train a triplet loss embedding to learn a 128 dimensional embedding optimized for verification and clustering. This network was trained using a private dataset of over 200M subjects. DeepFace [5][7] uses a deep network coupled with 3D alignment, to normalize facial pose by warping facial landmarks to a canonical position prior to encoding. DeepID2+ [9] and DeepID3 [8] extended the inception architecture to include joint Bayesian metric learning [20] and multi-task learning for both identification and verification.

These top performing convolutional network architectures have interesting common properties. First, they all exhibit deep convolutional network structure, often with parallel specialized sub-networks. However, Parkhi et. al [4] showed that the VGG-16 very deep architecture [18], when trained with a broad and deep dataset containing one thousand examples of 2622 subjects, outperformed networks with specialized networks [6] and ensembles [8] on YouTubeFaces. Second, many top performing approaches use some form of pose normalization such as 2D/3D alignment [5], [4], [16] to warp the facial landmarks into a canonical frontal pose. Finally, many approaches use metric learning in the form of triplet loss similarity or joint Bayesian metric learning for the final loss to learn an optimal embedding for verification [6], [4], [15]. A recent independent study reached a similar conclusion that ensembles and metric learning are crucial for strong performance on LFW [21].

Recent evaluations on IJB-A [3] are also based on convolutional networks and mirror the top performing approaches on LFW and YouTubeFaces. Recent approaches include deep networks using triplet loss similarity[14][22] and joint

Bayesian metric learning [15], and five pose specialized sub-networks with 3D pose rendering [16]. Face-BCNN [17] applies the bilinear CNN architecture to face identification, publishing the earliest results on IJB-A.

Finally, we note that the approach of defining a similarity function for face verification using linear SVMs trained on a large negative set was originally proposed as one-shot similarity (OSS) [23][24], and applied in [25] for identification. We study the more general form of this concept, by considering templates of images and videos, alternative fusion strategies, and the impact of gallery negative sets for identification.

III. Template Adaptation

Template adaptation is a form of transfer learning, combining deep convolutional network features trained on a source domain of many labeled faces, with template specific linear SVMs trained on a target domain using the media in a template. Template adaptation can be further decomposed into probe adaptation for face verification, and probe and gallery adaptation for face identification.

First, we provide preliminary definitions. A media observation x is either a color image of a subject, or a set of m video frames of a subject. An image encoding $z = f(x)$ is a mapping $f(x) \in \mathcal{R}^d$ from an image x to an encoding z with dimensionality d (e.g. features from a deep CNN). An average encoding $\bar{z} = \frac{1}{m}\sum_x f(x)$ is the average of image/frame encodings in a media observation, such as the encodings for all frames in a video. A template X is a set of encoded media observations $X = \{f(x_1), f(x_2), \ldots, f(x_k)\}$ of one subject. The size of a template $|X|$ is defined as the number of unique media used for encoding. Finally, a gallery $G = \{(X_1, y_1), (X_2, y_2), \ldots, (X_m, y_m)\}$ is a set of tuples of templates X and associated subject identification label y.

Figure 1 shows an overview of this concept. Each colored shape corresponds to a feature encoding of image or a video feature for the media in a template, such as generated from a convolutional network trained offline. The gray squares correspond to encodings of a large set of media of subjects disjoint from the subject identities in the gallery. The centroid of the colored shapes corresponds to the average encoding for this template. Probe adaptation is the problem of max-margin classification of the positive features from a template to the large negative feature set. The similarity between the blue probe template and the mated (genuine subject) green template is the margin (dotted lines) of the green feature encodings to the decision surface. Observe that this margin is positive, whereas the margin for the red classifier is negative, so that the blue/green similarity is much larger than blue/red as desired. Gallery adaptation is the problem of max-margin classification where the negative feature set for the gallery templates are defined by the other gallery templates.

More formally, probe adaptation is the training of a similarity function $s(P, Q)$ for a probe template P and reference template Q. Train a linear SVM for P, using unit normalized average encodings of media in P as positive features and a large feature set as negatives. The large negative set contains one feature encoding for many subject identities, so this set is very likely to be disjoint with the probe template. Similarly, train a linear SVM for Q, using the unit normalized average encodings for media in Q as positive features and a large feature set as negatives. Finally, let $P(q)$ be notation for evaluating the SVM functional margin (e.g. $w^T x$) trained on P, and evaluated using the unit normalized average media encoding q in template Q. The final similarity score for probe adaptation is the fusion of the two classifier margins using a linear combination $s(P, Q) = \frac{1}{2}P(q) + \frac{1}{2}Q(p)$. For implementation details, see section IV-A.

Gallery adaptation is the training of a similarity function $s(P, G)$ from a probe template P to gallery G. A gallery contains templates $G = \{X_1, X_2, \ldots, X_m\}$, and gallery adaptation trains a linear SVM for all pairs $s(P, X_i)$ following the approach for probe adaptation. Gallery adaptation differs from probe adaptation in that the large negative set for a template X_i is all unit normalized media encodings from all other templates in G not including X_i. In other words, the other non-mated subjects (e.g. subjects with different identities) in the gallery are used to construct negative features for X_i, whereas the large negative set is used for P. The final similarity score for gallery adaptation is the fusion of the probe classifier and the gallery classifier for each $X \in G$ using the linear combination $s(P, X) = \frac{1}{2}P(x) + \frac{1}{2}X(p)$.

IV. Results

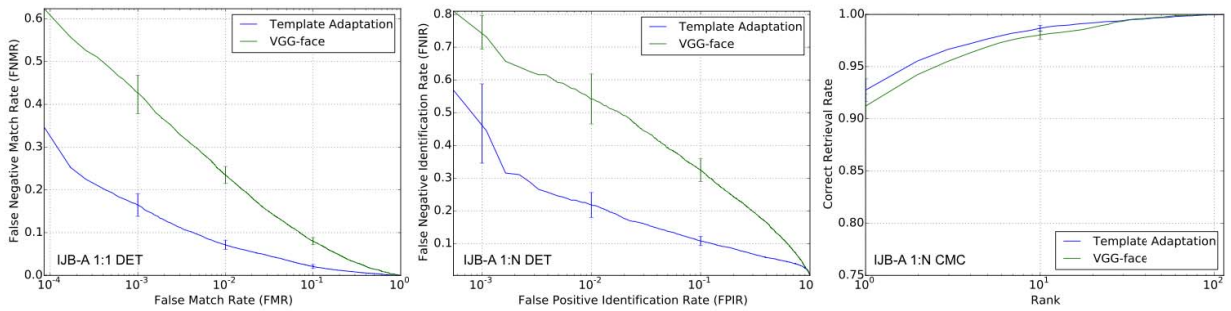
The proposed approach in section III introduces a number of research questions to study.

How does this compare to the state of the art? In section IV-B, we compare the template adaptation approach to all published results and show that the proposed approach exceeds the state of the art by a wide margin. Furthermore, in section IV-C we perform an analysis of alternatives to combine the state of the art techniques with template adaptation and show that when combined, these alternative approaches all result in nearly the same performance.

How should the negative set be formed? Template adaptation requires training linear SVMs, which require a labeled set of positive and negative feature encodings. In section IV-D, we perform a study to evaluate different strategies of constructing this negative set including using a holdout set, external negative set and combinations. Results show that the gallery based negative set is best for gallery adaptation.

How large do the templates need to be? In section IV-E, we study the effect of template size, or total number of media in a template, on verification performance to identify the minimum template size necessary, to help guide future template based dataset construction. We show that a minimum of three unique media per template results in diminishing returns for template adaptation.

What are the error modes of the template adaptation? In section IV-F, we visualize the best and worst templates pairs in IJB-A for verification, and we show that template size (e.g. number of media in a template) has the largest effect on performance.



IJB-A 1:1	FNMR@FMR=0.01
Template Adaptation	**0.061 ± 0.013**
VGG-Face [4]	0.195 ± 0.030
Joint Bayesian [15]	0.162 ± 0.042
Triplet Similarity [14]	0.21 ± 0.03
Deep Multipose [16]	0.213
Face-Search [26]	0.267 ± 0.034
Bilinear-CNN [17]	N/A

IJB-A 1:N	TPIR@FPIR=0.01	TPIR@FPIR=0.1	TPIR@Rank=1	TPIR@Rank=10
Template Adaptation	**0.774 ± 0.049**	**0.882 ± 0.016**	**0.928 ± 0.010**	**0.986 ± 0.003**
VGG-Face [4]	0.461 ± 0.077	0.670 ± 0.031	0.913 ± 0.011	0.981 ± 0.005
Joint Bayesian [15]	0.5768 ± 0.0940	0.7897 ± 0.0333	0.903 ± 0.012	0.977 ± 0.007
Triplet Similarity [14]	0.5560 ± 0.0654	0.7541 ± 0.0141	0.8801 ± 0.0148	0.9738 ± 0.0057
Deep Multipose [16]	0.52	0.75	0.846	0.947
Face-Search [26]	0.383 ± 0.063	0.613 ± 0.032	0.820 ± 0.024	N/A
Bilinear-CNN [17]	0.142 ± 0.027	0.341 ± 0.032	0.588 ± 0.022	N/A

Fig. 2. IJB-A Evaluation. (top) 1:1 DET for verification (lower is better), 1:N DET for identification (lower is better) and CMC for identification (higher is better) shown for template adaptation and VGG-face [4]. (bottom) Performance at operating points as compared to published results sorted by rank-1 recall (true positive identification rate or TPIR) for VGG-face [4], Bilinear-CNN [17], Joint Bayesian [15], Triplet Similarity [14], Face-Search [26] and Deep Multipose [16]. Results show that Template Adaptation sets a new state-of-the-art by a wide margin.

How should template classifier scores be fused? We study the effect of different strategies for combination of two classifiers, based on winner take all and weighted combinations based on template size. We conclude that an average combination is the best fusion approach [27].

A. Experimental System

We use the VGG-Face deep convolutional neural network [4], using the penultimate layer output as the feature encoding f. For computing the average encoding across frames of video, we use *face tracks* which compute the mean encoding of all frames in a video followed by unit normalization. This approach was shown to be effective for Fisher vector encoding [28] and deep CNN encoding [4].

Media encoding is preprocessed according to the following pipeline. For each media, we crop each face using the ground truth or detected facial bounding box dilated by a factor of 1.1. Then, we anisotropically rescale this face crop to 224x224x3, such that the aspect ratio may not be preserved. This is the assumed input size for the CNN. Next, we encode this face crop for each image or frame in the template using the VGG-face network, and compute average video encodings for each video. Next, we unit normalize each media feature, and train the weights and bias for a linear SVM for each template. We use the LIBLINEAR library with L2-regularized L2-loss primal SVM with class weighted squared hinge loss objective [29]. Cross validation experiments showed that this squared hinge loss performed better than the more common L1-loss.

$$\min_w \frac{1}{2} w^T w + C_p \sum_{i=1}^{N_p} \max[0, 1 - y_i w^T x_i]^2 + C_n \sum_{j=1}^{N_n} \max[0, 1 - y_j w^T x_j]^2 \quad (1)$$

The loss in (1) includes terms for both positive and negative features, such that C_p is the regularization constant for N_p positive observations ($y_i = +1$) and C_n for negative observations ($y_i = -1$). This formulation of the loss enables data rebalancing for cases where $N_p << N_n$. The positive features in N_p are the average media encodings in the template. The negative features are derived from a large negative feature set in N_n (either from a large negative set for probe adaptation, or other non-mated templates for gallery adaptation). The parameters $C_p = C\frac{N_p+N_n}{2N_p}$ and $C_n = C\frac{N_p+N_n}{2N_n}$ adjust the regularization constants to be proportional to the inverse class frequency. The parameter $C = 10$ in the SVM, trading-off regularizer and loss, was determined using an held-out validation subset of the data. Finally, the learned weights w include a bias term by augmenting x with a constant dimension of one.

At test time, we evaluate the linear SVMs as described in section III. We compute the average media encodings for each media in a template, then compute the mean of the media encodings, then unit normalize forming a template encoding. This constructs a single feature for each template. Given two templates P and Q, let $P(q)$ be the evaluation of the functional SVM margin (e.g. $P(x) = w^T x$) for the trained linear SVM for P, given the template encoding q

for Q. Finally, the similarity $s(P,Q) = \frac{1}{2}P(q) + \frac{1}{2}Q(p)$ is a weighted combination of the functional margins for the SVM for P evaluated on encoding q and Q evaluated on p.

For baseline comparison, we use the VGG-face network with 4096d features encoded from the penultimate fully connected layer. Media encodings are constructed by averaging features across a video [28], [4], and template encodings are constructed by averaging media encodings over a template, then unit normalizing. Template similarity is equivalent to negative L2 distance over unit normalized template encodings. We also compare results with 2D alignment, triplet similarity embedding and joint Bayesian triplet similarity embedding. For the triplet loss and joint Bayesian metric learning, we use hyperparameter settings such that minibatch = 1800, 1M semi-hard [6] negative triplets per minibatch, dropconnect regularization [30], 3 epochs of Parallel SGD [31], fixed learning rate $\nu = 0.25$. For 2D alignment, we use ground truth facial bounding boxes and facial landmark regression [32], followed by a robust least squares similarity transform estimation to best center the nose.

For all research studies in sections IV-C - IV-F, we report 1:1 verification ROC curve for all probe and gallery template pairs in IJB-A split 1 and CMC for identification on IJB-A split 1 (see section IV-B for definitions). This is equivalent to IARPA Janus Challenge Set 2 (CS2) evaluation protocol, which is also reported in the literature.

Finally, we analyze the runtime impacts on training probe and gallery adaptation as compared to a deep CNN only. Figure 3 shows timing results for template flattening, which includes computing the template encodings and linear SVM training for each template in IJB-A split 1. These CNN encodings were performed on an NVIDIA Tesla K40, and SVM training was performed on 1.8GHz Intel Xeon E5-2650L, with LIBLINEAR multithreading disabled. We report the time for probe flattening, which is encoding and SVM training probe templates, and gallery flattening which is encoding and SVM training for gallery templates. Also, we report the time for verification and search, which includes computing similarity scores for pairs of templates. Results show that template adaptation is slower by a factor of 5.6x for probe adaptation, 5.9x for gallery flattening, 2.2x slower for search and 2.0x slower for verification. Experiments show that performance can be further improved by 2.2x using multi-threaded liblinear support.

Timing Evaluation (sec)	Probe Adaptation	Gallery Adaptation	Verification	Search
Template Adaptation	3307.0	1154.2	25.9	20.9
VGG-face	592.2	197.0	12.8	9.7

Fig. 3. Timing Analysis. GPU optimized template adaptation is slower by a factor of 2.2x for search and 2.0x for verification than CNN only, and probe flattening is slower by a factor 5.6x, gallery flattening by 5.9x.

B. IJB-A Dataset and Evaluation

In this section, we describe the results for evaluation of the experimental system on the IJB-A verification and identification protocols [3]. IJB-A contains 5712 images and 2085 videos of 500 subjects, for an average of 11.4 images and 4.2 videos per subject. This dataset was manually curated using Mechanical Turk from media-in-the-wild to annotate the facial bounding box and eyes and nose facial landmarks, and this manual annotation avoids the Viola-Jones near-frontal bias. Furthermore, this dataset was curated to control for ethnicity, country of origin and pose biases.

Metrics for 1:1 verification are evaluated using a decision error tradeoff (DET) curve. The 1:1 DET curve is equivalent to a receiver operating characteristics (ROC) curve, where the true accept rate is one minus the false negative match rate. This evaluation plots the false negative match rate vs. the false match rate as a function of similarity threshold for a given set of pairs of templates for verification.

Metrics for 1:N identification are the Decision Error Tradeoff (DET) curve and the Cumulative Match Characteristic (CMC) curve. The 1:N DET curve plots the false negative identification rate vs. the false positive identification rate as a function of similarity threshold for a search of L=20 candidate identities in a gallery. The 1:N CMC curve is an information retrieval metric that captures the recall of a specific probe identify within the top-K most similar candidates when searching the gallery. This DET curve is appropriate for limiting the workload for an analyst by allowing for a similarity threshold to be applied to reject false matches even if in the top-K. For detailed description of these metrics, refer to [3], [13].

Performance evaluation for IJB-A requires evaluation of ten random splits of the dataset into training and testing (gallery and probe) sets. The evaluation protocol for 1:1 verification considers specific pairs of mated (genuine) and non-mated (imposter) subjects. The non-mated pairs were chosen to control for gender and skin tone to make the verification problem more challenging. Performance is reported for operating points on each of the curves: 1:1 DET reports false negative match rate at a false match rate of 1e-2, 1:N DET report true positive identification rate (e.g. 1-false negative identification rate) at false positive identification rate of 1e-2, and CMC report true positive identification rate (recall or correct retrieval rate) at rank-one and rank-ten. The 10 splits are used to compute standard deviations for each of these operating points, to characterize statistical significance.

Figure 2 shows the overall evaluation results on IJB-A. This evaluation compares the baseline approach of VGG-Face only using the fc7 embedding from training a face classifier as described in [4] with the proposed approach of VGG-Face encoding with probe and gallery template adaptation. These results show that identification performance is slightly improved for rank 1 and rank 10 retrieval, however there are large performance improvements for the 1:N DET for identification and the 1:1 DET for verification. The table in figure 2 shows performance at specific operating points for verification and identification, and compares to published results in the literature for joint Bayesian metric learning [15], triplet similarity embedding [14], multi-pose learning [16], bilinear CNNs [17] and very deep CNNs

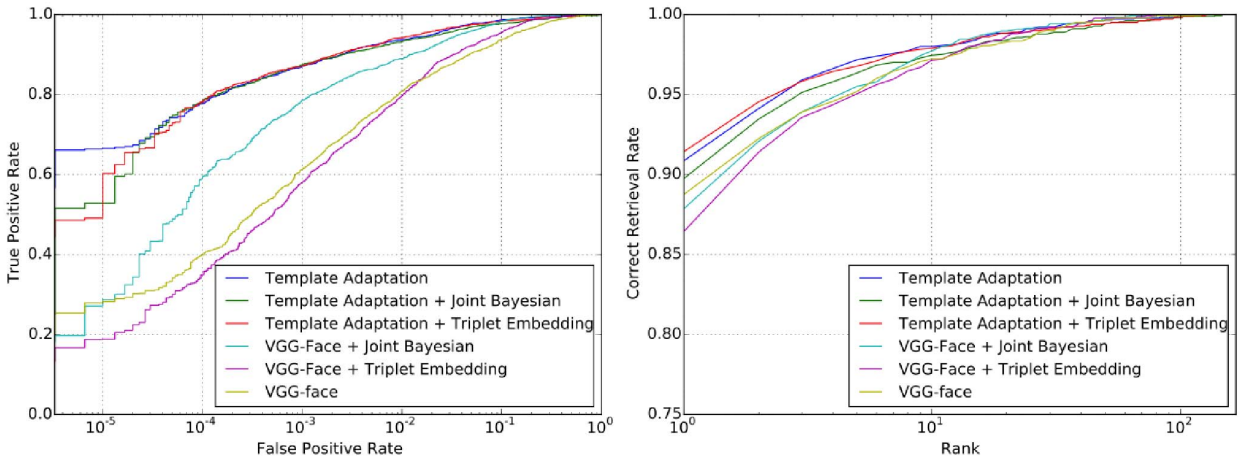


Fig. 4. Analysis of Alternatives. We show verification ROC curves (left) and identification CMC curves (right) for IJB-A split-1. (top) Template adaptation compared with CNN encoding with metric learning using triplet similarity embedding [4], [6] or Joint Bayesian embedding [20], [22]. Template adaptation compared with CNN encoding and 2D alignment [5], [4] is shown [27]. In both cases, template adaptation outperforms all methods, and when combined with metric learning or 2D alignment, generates nearly equivalent performance.

[4], [26]. These results show that the proposed template adaptation, while conceptually simple, exhibits state-of-the-art performance by a wide margin on this dataset.

C. Analysis of Alternatives

Figure 4 shows an analysis of alternatives study. The state of the art approaches on LFW and YouTubeFaces often augment a very deep CNN encoding with metric learning [6], [4] for improved verification scores or 2D alignment [5], [4] to better align facial bounding boxes. In this study, we implement triplet loss similarity embedding, joint Bayesian similarity embedding and 2D alignment [27], and use these alternative feature encodings as input to template adaptation.

We report 1:1 DET for all probe and gallery template pairs in IJB-A split 1 and CMC for identification on IJB-A split 1. This study shows that template adaptation on the CNN output provides nearly the same result as template adaptation with metric learning or 2D alignment based features. This implies that the additional training and computational requirements for these approaches are not necessary for template based datasets. Furthermore, we show that 2D alignment does not provide much benefit on IJB-A [27], in contrast with reported performance on near frontal datasets [4], [5]. One hypothesis is that this dataset includes many profile faces for which facial landmark alignment is inaccurate or fails altogether.

D. Negative Set Study

We study the effect of different combinations of negative feature sets on overall verification performance. Recall that probe and gallery template adaptation require the use of a large negative set for training each linear SVM. This study compares using combinations of features drawn from the non-mated subjects in the gallery (neg) and features drawn from an independent subject disjoint training set (trn). This training set is drawn from the same dataset distribution as the gallery, but is subject disjoint. This dataset distribution was constructed by matching the mean number of unique videos and images per subject in the the gallery.

Additional results provided in supplementary material [27] show that using the gallery set as negative feature set provides the best performance for gallery adaptation. Using the disjoint training set for probe adaptation is the best for verification. This is the final strategy used for evaluation in figure 2. This conclusion is somewhat surprising that the probe adaptation was worse when constructing a negative set combining neg+trn, as a larger negative set typically results in better generalization performance for related approaches such as exemplar-SVM [33]. However, a larger negative set would dilute the effect of the discriminating between gallery subjects, which is the primary goal of the evaluation, so a focused negative set would be appropriate. Finally, in the supplementary material [27], we also perform a negative set analysis using the CASIA WebFaces dataset [34], showing only slightly reduced verification performance.

E. Template Size Study

Figure 5 shows an analysis of performance as a function of template size. For this study, we consider pairs of templates (P, Q) and compute the maximum template size as $\max(|P|, |Q|)$. Next, we consider max template sizes in the range $(1, 2), (2, 4), (4, 8), (8, 16), (16, 32)$ and $(32, 64)$, and compute a verification ROC curve for only those template pairs with sizes within the range. For each, we report a single point on the ROC curve at a false alarm rate of 1e-2 or 1e-3. Results from section IV-B show that the largest benefit for template adaptation is on verification performance, so we analyze the effect of the template sizes on this metric.

Figure 5 (left) shows mean similarity score for templates of mated subjects within a given template size range. This shows that as the template size increases the mated similarity score also increases. This is perhaps not surprising, as the more observations of media that are available in a template,

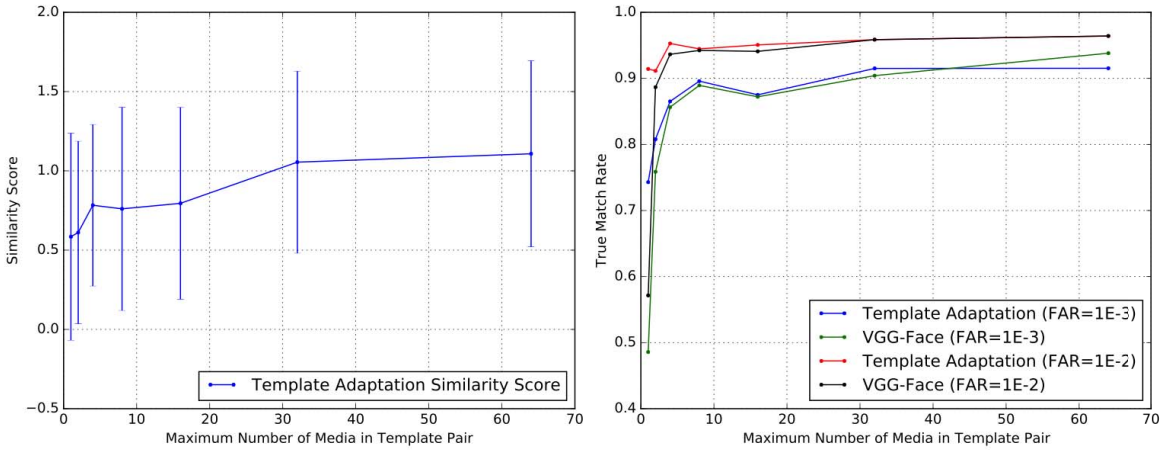


Fig. 5. Template size analysis. (left) Similarity score increases as a function of maximum number of media, where the standard deviation is largest when template size is at least one, although not by a significant amount (right) True match rate as a function of maximum number of unique images or videos in a template pair, which shows that verification performance levels off at a maximum of *three* unique media per template.

the better the subject representation and the better the similarity score. The largest uncertainty as shown by the error bars is when the maximum template size is one, which is also not too surprising. Interestingly, the similarity score variance does not decrease as template sizes increase, rather they stay largely the same even as the mean similarity increases.

Figure 5 (right) shows the effect of template size on verification performance. For each point on this curve, we split the dataset into templates that contained sizes within the range shown. Then, we computed a ROC curve and report the true match rate at a false alarm rate of 1e-3 and 1e-2, an operating point on the verification ROC curve. This result shows that the rate of increase in performance is largest for few media, and performance saturates at about 3 media per template. Furthermore, as the number of media per template increases, the verification score at 1e-2 increases by about 19% from one media per template to sixty four. This also shows that the largest benefit for template adaptation is when there are at least three media per template.

F. Error Analysis

Finally, we visualized identification and verification errors in different performance domains, in order to gain insight into template-based facial recognition. More detailed figures and additional discussion, including identification analysis, are available in supplementary material [27].

Figure 6 shows four columns of verification probe and gallery pairs for: the best scoring mated pairs; worst scoring mated pairs; best scoring non-mated pairs; and worst scoring non-mated pairs. Figure 6 (a) shows the highest mated similarities. In the thirty highest scoring correct matches, we immediately note that every gallery template contains dozens of media. The probe templates either contain dozens of media or one media that matches well. Figure 6 (b) shows the lowest mated template pairs, representing failed identification. The thirty lowest mated similarities result from single-media probe templates that are low contrast, low resolution, extremely non-frontal, or not oriented upwards.

Figure 6 (c) shows the worst non-mated pairs, which highlights understandable errors involving single-media probe templates representing impostors in challenging orientations. Figure 6 (d) shows the best non-mated similarities, which often involve large templates.

V. Conclusions

In this paper, we have studied and extended template adaptation, a simple and surprisingly effective strategy for face verification and identification that achieves state of the art performance on the IJB-A dataset. Furthermore, we showed that this strategy can be applied to existing networks to improve performance. Futhermore, our evaluation provides compelling evidence that there are many face recognition tasks that can benefit from a historical record of media to aid in matching, and that this is an important problem to further evaluate with new template-based face datasets.

Our analysis shows that performance is highly dependent on the number of media available in a template. This strategy results in performance that results in 19% decrease in verification scores when a template contains a single media, such as comparing image to image or video to video, as in LFW or YouTubeFaces style evaluations. However, when probe or gallery templates are rich and at least one template contains greater than three media, performance quickly saturates and dominates the state of the art.

References

[1] Huang, G., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: University of Massachusetts, Amherst, Technical Report 07-49. (2007)

[2] Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: CVPR. (2011)

Fig. 6. Verification error analysis. (a) The best mated verification template pairs, (b) The worst mated verification template pairs, (c) The worst non-mated verification template pairs (d) The best non-mated verification template pairs.

[3] Klare, B., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.: Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In: CVPR. (2015)

[4] Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. (2015)

[5] Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: Closing the gap to human-level performance in face verification. In: CVPR. (2014)

[6] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: CVPR. (2015)

[7] Y. Taigman, M. Yang, M.R., Wolf, L.: Web-scale training for face identification. In: CVPR. (2015)

[8] Sun, Y., Liang, D., Wang, X., Tang., X.: DeepID3: Face recognition with very deep neural networks. In: arXiv:1502.00873. (2014)

[9] Sun, Y., Wang, X., Tang., X.: Deeply learned face representations are sparse, selective, and robust. In: CVPR. (2015)

[10] Learned-Miller, E., Huang, G., RoyChowdhury, A., Li, H., Hua, G.: Labeled Faces in the Wild: A Survey. In: Advances in Face Detection and Facial Image Analysis. Springer (2015)

[11] Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with GaussianFace. In: AAAI. (2015)

[12] Phillips, J., Hill, M., Swindle, J., O'Toole, A.: Human and algorithm performance on the pasc face recognition challenge. In: BTAS. (2015)

[13] Grother, P., Ngan, M.: Face recognition vendor test (frvt): Performance of face identification algorithms. In: NIST Interagency Report 8009. (2014)

[14] Sankaranarayanan, S., Alavi, A., Chellappa, R.: Triplet similarity embedding for face verification. In: arXiv:1602.03418. (2016)

[15] J. Chen, V.P., Chellappa, R.: Unconstrained face verification using deep CNN features. In: WACV. (2016)

[16] AbdAlmageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., Choi, J., Lekust, J., Kim, J., Natarajan, P., Nevatia, R., Medioni, G.: Face recognition using deep multi-pose representations. In: WACV. (2016)

[17] RoyChowdry, A., Lin, T., Maji, S., Learned-Miller, E.: One-to-many face recognition with bilinear CNNs. In: WACV. (2016)

[18] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)

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

[20] Chen, D., Cao, X., Wang, L., Wen, F., Sun, J.: Bayesian face revisited: A joint formulation. In: ECCV. (2012)

[21] Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S.Z., Hospedales, T.: When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In: ICCV workshop on ChaLearn Looking at People. (2015)

[22] Chen, J., Ranjan, R., Kumar, A., Chen, C., Patel, V., Chellappa, R.: An end-to-end system for unconstrained face verification with deep convolutional neural networks. In: ICCV workshop on ChaLearn Looking at People. (2015)

[23] Wolf, L., Hassner, T., Taigman, Y.: The one-shot similarity kernel. In: ICCV. (2009)

[24] Wolf, L., Hassner, T., Taigman, Y.: Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. PAMI **33**(10) (2011)

[25] Chatfield, K., Arandjelović, R., Parkhi, O.M., Zisserman, A.: On-the-fly learning for visual search of large-scale image and video datasets. International Journal of Multimedia Information Retrieval (2015)

[26] Wang, D., Otto, C., Jain, A.: Face search at scale: 80 million gallery. In: arXiv:1507.07242. (2015)

[27] Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., Zisserman, A.: Template adaptation for face verificaiton and identification. In: arXiv:1603.03958v3. (2016)

[28] Parkhi, O.M., Simonyan, K., Vedaldi, A., Zisserman, A.: A compact and discriminative face track descriptor. In: CVPR. (2014)

[29] Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: Liblinear: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874

[30] Wan, L., Zeiler, M., Zhang, S., LeCun, Y., Fergus, R.: Regularization of neural network using dropconnect. In: ICML. (2013)

[31] Zinkevich, M., et al: Parallelized stochastic gradient descent. In: NIPS. (2011)

[32] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR. (2014)

[33] Malisiewicz, T., Gupta, A., Efros, A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV. (2011)

[34] Yi, D., Lei, Z., Liao, S., Li, S.: Learning face representation from scratch. In: arXiv:1411.7923. (2014)