Dependency-aware Attention Control for Unconstrained Face Recognition with Image Sets

Xiaofeng Liu[1][0000−0002−4514−2016], B.V.K Vijaya Kumar[1][0000−0001−7126−6381], Chao Yang[2][0000−0002−6553−7963], Qingming Tang[3][0000−−0002−2670−4917], and Jane You[4][0000−0002−8181−4836]

[1] Carnegie Mellon University PA 15213, USA
liuxiaofeng@cmu.edu
[2] University of Southern California CA 90089, USA
[3] Toyota Technological Institute at Chicago IL 60637, USA
[4] The Hong Kong Polytechnic University

Abstract. This paper targets the problem of image set-based face verification and identification. Unlike traditional single media (an image or video) setting, we encounter a set of heterogeneous contents containing orderless images and videos. The importance of each image is usually considered either equal or based on their independent quality assessment. How to model the relationship of orderless images within a set remains a challenge. We address this problem by formulating it as a Markov Decision Process (MDP) in the latent space. Specifically, we first present a dependency-aware attention control (DAC) network, which resorts to actor-critic reinforcement learning for sequential attention decision of each image embedding to fully exploit the rich correlation cues among the unordered images. Moreover, we introduce its sample-efficient variant with off-policy experience replay to speed up the learning process. The pose-guided representation scheme can further boost the performance at the extremes of the pose variation.

Keywords: Deep Reinforcement Learning · Actor-Critic · Face recognition · Set-to-Set · Attention Control

1 Introduction

Recently, unconstrained face recognition (FR) has been rigorously researched in computer vision community [1, 2]. In its initial days, the single image setting is used for FR evaluations, *e.g.*, Labeled Faces in the Wild (LFW) verification task [3]. The trend of visual media explosion pushes the research into the next phase, where the video face verification attracts much attention, such as the YouTube Faces (YTF) dataset [4]. Since the LFW and YTF have a well-known frontal pose selection bias, the unconstrained FR is still considered an unsolved problem [5, 6]. In addition, the open-set face identification is actually more challenging compared to the verification popularized by the LFW and YTF datasets [7, 8].

The IARPA Janus Benchmark A (IJB-A) [9] provides a more practical unconstrained face verification and identification benchmark. It takes a set (containing

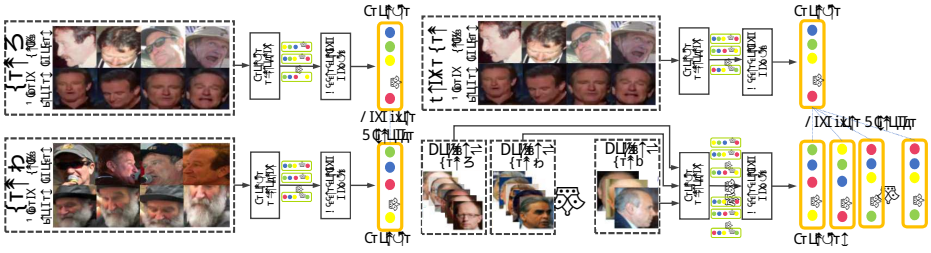


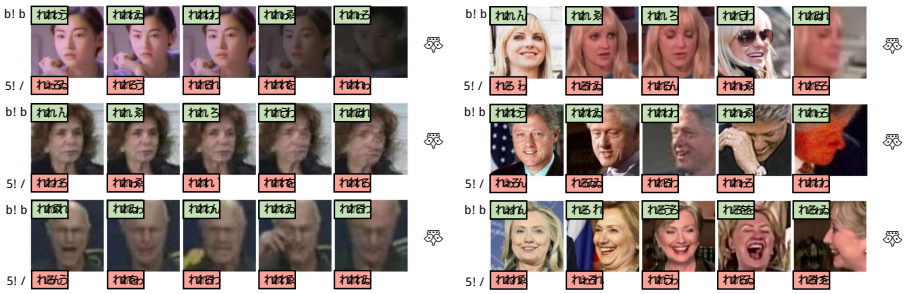
Fig. 1. Illustration of the image set-based 1:1 face verification (left) and open-set 1:N identification (right) using the typical aggregation method.

orderless images and/or videos with extreme head rotations, complex expressions and illuminations) as the smallest unit for representation. The set of a subject can be sampled from the mugshot history of a criminal, lifetime enrollment images for identity documents, different check points, and trajectory of a face in the video. This kind of setting is more similar to the real-world biometric scenarios [10]. Capturing human faces from multiple views, background environments, camera parameters, does result in the large inner-set variations, but also incorporates more complementary information hopefully leading to higher accuracy in practical applications [11].

A commonly adopted strategy to aggregate identity information in each image is the average/max pooling [12–15]. Since the images vary in quality, a neural network-based assessment module has been introduced to independently assign the weight for each frame [11, 16]. By doing this, the frontal and clear faces are favored by their model. However, this may result in redundancy and sacrifice the diversity in a set. As shown in Fig. 2, these inferior frontal images are given relatively high weights in a set, sometimes as high as the weight given to the most discriminative one. There is little additional information that can be extracted from the blurry version of the same pose, while the valuable profile information *etc.*, are almost ignored by the system. We argue that the desired weighting decision should depends on the other images within a set.

Instead, we propose to formulate the attention scheme as a Markov Decision Process and resort to the actor-critic reinforcement learning (RL) to harness model learning. The dependency-aware attention control (DAC) module learns a policy to decide the importance of each image step-by-step with the observation of the other images in a set. In this way, we adaptively aggregate the feature vectors into a highly-compact representation inside the convex hull spanned by them. It not only explicitly learns to advocate high-quality images while repelling low-quality one, but also considers the inner-set dependency to reduce the redundancy and maintains the benefit of diversity information.

Moreover, extracting a set-level invariable feature can be always challenging to incorporate all of the potential information in varying poses, illumination conditions, resolutions *etc.* Some approaches aggregate the image-level pair-wise similarity scores of two compared sets to fully use all images [17–20]. Given n as



(a) (b)

Fig. 2. Typical examples on the test set of (a) YTF and (b) IJB-A dataset showing the weights of images calculated by the previous method NAN [16], and proposed DAC.

the average number of images in a set, then this corresponds to the $\mathcal{O}(n^2)$ computational complexity per match operation and $\mathcal{O}(n)$ space complexity per set are not desirable. More recently, [21–23] are proposed to trade-off between speed and accuracy of processing paired video inputs using value-based Q-learning methods. These configurations focus on the verification, and cannot be scaled well for large scale identification tasks [16]. Conventionally, the feature extraction of the probe and gallery samples are independent processes [8].

We notice that pose is the primary challenge in the IJB-A dataset and real applications [7, 24, 25], and there is a prior that the structures of frontal and profile face are significantly different. Therefore, we simply utilize a pose-guided representation (PGR) scheme with stochastic routing to model the inter-set dependency. It well balances the computation cost and information utilization.

Considering the above factors, we propose to fully exploit both the inner- and inter-set relationship for the unified set-based face verification and identification. (1) To the best of our knowledge, this is the first effort to introduce deep actor-critic RL into visual recognition problem. (2) The DAC can potentially be a general solution to incorporate rich correlation cues among orderless images. Its coefficients can be trained in a normal recognition training task given only set-level identity annotation, without the need of extra supervision signals. (3) To further improve the sample-efficiency, the trust region-based experience replay is introduced to speedup the training and achieve a stronger convergence property. (4) The PGR scheme well balances the computation cost and information utilization at the extremes of pose variation with the prior domain knowledge of human face. (5) The module-based feature-level aggregation also inherits the advantage of conventional pooling strategies *e.g.*, taking varied number of inputs as well as offering time and memory efficiency.

We show that our method leads to the state-of-the-art accuracy on IJB-A dataset and also generalizes well in several video-based face recognition tasks, *e.g.*, YTF and Celebrity-1000.

2 Related work

Image set/video-based face recognition has been actively studied in recent years [2]. The multi-image setting in the template-based dataset is similar to the multiple frames in the video-base recognition task. However, the temporal structure within a set is usually disordered, and the inner/inter-set variations are more challenging [9]. We will not cover the methods which exploit the temporal dynamics here. There are two kinds of conventional solutions, *i.e.*, manifold-based and image-based methods. In the first category, each set/video is usually modeled as a manifold, and the similarity or distance is measured in the manifold-level. In previous works, the affine hull, SPD model, Grassmann manifolds, n-order statistics and hyperplane similarity have been proposed to describe the manifolds [26–30]. In these methods, images are considered as equal importance. They usually cannot handle the large appearance variations in the unconstrained FR task. For the second category, the pairwise similarities between probe and gallery images are exploited for verification [17, 18, 20, 31, 32]. The quadratic number of comparisons make them not scale well for identification tasks. Yang *et al.* [16] propose an attention model to aggregate a set of features to a single representation with an independent quality assessment module for each feature. Reference [33] further up-sampled the aggregated features to an image, then fed it to an image-based FR network. However, weighting decision for an image does not take the other images into account as discussed in Sec. 1. Since the frequently used RNN in video task [34, 35, 22] is not fit for the image set, in this work, we consider the dependency within a set of features in a different way, where we use deep reinforcement learning to suggest the attention of each feature.

Reinforcement learning (RL) trains an agent to interact (by trail and error) with a dynamic environment with the objective to maximize its accumulated reward. Recently, deep RL with convolutional neural networks (CNN) achieved human-level performance in Atari Games [36]. The CNN is an ideal approximate function to address the infinite state space [37]. There are two main streams to solve RL problems: methods based on value function and methods based on policy gradient. The first category, *e.g.*, Q-learning, is the common solution for discrete action tasks [36]. The second category can be efficient for continuous action space [38, 39]. There is also a hybrid actor-critic approach in which the parameterized policy is called an actor, and the learned value-function is called a critic [40, 41]. As it is essentially a policy gradient method, it can also be used for continuous action space [42].

Besides, policy-based and actor-critic methods have faster convergence characteristics than value-based methods [43], but they usually suffer from low sample-efficiency, high variance and often converge to local optima, since they typically learn via on-policy algorithms [44, 45]. Even the Asynchronous Advantage Actor-Critic [40, 41] also requires new samples to be collected for each gradient step on the policy. This quickly becomes extravagantly expensive, as the number of gradient steps to learn an effective policy increases with task complexity. Off-policy learning instead aims to reuse past experiences. This is

not directly feasible with conventional policy gradient formulations, despite it relatively straightforward for value-based methods [37]. Hence in this paper, we focus on combining the stability of actor-critic methods with the efficiency of off-policy RL, which capitalizes in recent advances on deep RL [40], especially off-policy algorithms [46, 47].

In addition to its traditional applications in robotics and control, recently RL has been successfully applied to a few visual recognition tasks. Mnih *et al.* [48] introduce the recurrent attention model to focus on selected regions or locations from an image for digits detection and classification. This idea is extended to identity alignment by iteratively removing irrelevant pixels in each image [49]. The value-based Q-learning methods are used for object tracking [50] and the video verification in a computationally efficient view by dropping inefficient probe-gallery pairs [22] or stopping the comparison after receiving sufficient pairs [21, 23] . However, this will inevitably result in information loss of the unused pairs and only applicable for verification. There has been little progress made in policy gradient/actor-critic RL for visual recognition.

3 Proposed methods

The flow chart of our framework is illustrated in Fig. 3. It takes a set of face images as input and processes them with two major modules to output a single(w/o PGR)/three(with PGR) feature vectors as its representation for recognition. We adopt a modern CNN module to embed an image into a latent space, which can largely reduce the computation costs and offer a practicable state space for RL. Then, we cascade the DAC, which works as an attention model reads all feature vectors and linearly combines them with adaptive weighting at the feature-level. Following the memory attention mechanism described in [34, 35, 16], the features are treated as the memory and the feature weighting is cast as a memory addressing procedure.These two modules can be trained in a one-by-one or end-to-end manner. We choose the first option, which makes our system benefit from the sufficient training data of the image-based FR datasets. The PGR scheme can further utilize the prior knowledge of human face to address a set with large pose variants.

3.1 Inner-set dependency control

In the set-based recognition task, we are given M sets/videos $(\mathcal{X}^m, y^m)_{m=1}^M$, where $\mathcal{X}^m$ is a image set/video sequence with varying number of images T^m (*i.e.,* $\mathcal{X}^m = \{x_1^m, x_2^m, \cdots, x_{T^m}^m\}$, x_t^m is the t-th image in a set) and the y^m is the corresponding set-level identity label. We feed each image x_t^m to our model, and its corresponding feature representation f_t^m are extracted using our neural embedding network. Here, we adopt the GoogLeNet [51] with Batch Normalization [52] to produce a 128-dimensional feature as our encoding of each image. With a relatively simple architecture, GoogLeNet has shown superior performance on several FR benchmarks. It can be easily replaced by other advanced

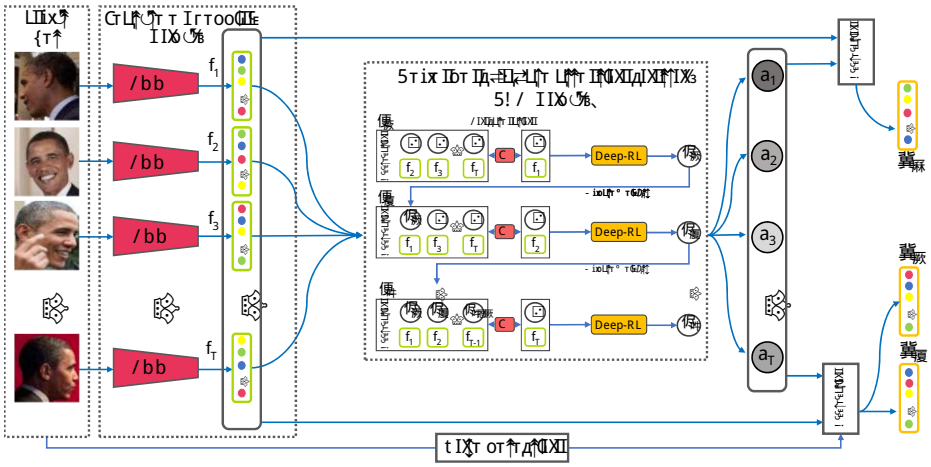


Fig. 3. Our network architecture for image set-based face recognition.

CNNs for better performance. In the rest of the paper, we will simply refer to our neural embedding network as CNN, and omit the upper index (identity) where appropriate for better readability.

Since the features are deterministically computed from the images, they also inherit and display large variations. Simply discarding some of them using the hard attention scheme may result in loss of too much information in a set [16, 22]. Our attention control can be seen as the task of reinforcement learning to find the optimal weights of soft attention, which defines how much of them are focused by the memory attention mechanism. Moreover, the principle of taking different number of images without temporal information, and having trainable parameters through standard recognition training are fully considered.

Our solution of inner-set dependency modeling is to formulated as a MDP. At each time step t, the agent receives a state s_t in a state space $\mathcal{S}$ and chooses an action a_t from an action space $\mathcal{A}$, following a policy $\pi(a_t \mid s_t)$, which is the behavior of the agent. Then the action will determine the next state *i.e.,* s_{t+1} or termination, and receive a reward $r_t(s_t, a_t) \in \mathcal{R} \subseteq \mathbb{R}$ from the environment. The goal is to find an optimal policy π^* that maximizes the discounted total return $R_t = \sum_{i \geq 0}^{T} \gamma^i r_{t+i}(s_t, a_t)$ in expectation, where $\gamma \in [0, 1)$ is the discount factor to trade-off the importance of immediate and future rewards [37].

In the context of image-set based face recognition, we define the actions, *i.e.,* $\{a^1, a^2, \cdots, a^T\}$, as the weights of each feature representation $\{f\}_{i=1}^{T}$. The weights of soft attention $\{a\}_{i=1}^{T}$ are initialized to be 1, and are updated step-by-step. The state s_t is related to the $t-1$ weighted features and $T-(t-1)$ to-be weighted features. In contrast to image-level dependency modeling, the compact embeddings largely shrink the state space and make our RL training feasible. In our practical applications, s_t is the concatenation of f_t and the aggregation of the remaining features with their updated weights at time step t. The termination

means all of the images in this set have been successfully traversed.

$$s_t = \left\{ \frac{(\sum_{i=1}^{T} a_i f_i) - f_t}{(\sum_{i=1}^{T} a_i) - 1} \right\} Concatenate\,\{f_t\} \tag{1}$$

We define the global reward for RL by the overall recognition performance of the aggregated embeddings, which drives the RL network optimization. In practice, we add on the top of the DAC few fully connected layers h and followed by a softmax to calculate the cross-entropy loss $L_m = -\log\left(\{e^{o_{y^m}}\}/\sum_j^M e^{o_j}\right)$ to calculate the reward at this time step. We use the notation o_j to mean the j-th element of the vector of class scores o. $g(\cdot)$ is the weighted average aggregation function, h maps the aggregated feature with the updated weights $g(\mathcal{X}^m \mid s_t)$ to the o. The reward are defined as follows:

$$g(\mathcal{X}^m \mid s_t, \mathrm{CNN}) = \sum_{i=1}^{T^m} \frac{a_i {f_i}^m}{\sum a_i} \quad \text{(with updated } a_i \text{ at step } t) \tag{2}$$

$$r_t = \{L_m[h(g(\mathcal{X}^m \mid s_t))] - L_m[h(g(\mathcal{X}^m \mid s_{t+1}))]\} + \lambda \max[0, (1 - a_t)] \tag{3}$$

where the hinge loss term serves as a regularization to encourage redundancy elimination and is balanced with the λ. It also contributes to stabilize training. The aggregation operation essentially selects a point inside of the convex hull spanned by all feature vectors [26].

Considering that the action space here is a continuous space $\mathcal{A} \in \mathbb{R}^+$, the value-based RL (*e.g.*, Q-Learning) cannot tackle this task. We adapt the actor-critic network to directly grade each feature dependent on the observation of the other features. In a policy-based method, the training objective is to find a parametrized policy $\pi_\theta(a_t \mid s_t)$ that maximizes the expected reward $J(\theta)$ over all possible aggregation trajectories given a starting state. Following the Policy Gradient Theorem [43], the gradient of the parameters given the objective function has the form:

$$\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_t \mid s_t)(Q(s_t, a_t) - b(s_t))] \tag{4}$$

where $Q(s_t, a_t) = \mathbb{E}[R_t \mid s_t, a_t]$ is the state-action value function, in which the initial action a_t is provided to calculate the expected return when starting in the state s_t. A baseline function $b(s_t)$ is typically subtracted to reduce the variance while not changing the estimated gradient [44, 53]. A natural candidate for this baseline is the state only value function $V(s_t) = \mathbb{E}[R_t \mid s_t]$, which is similar to $Q(s_t, a_t)$, except the a_t is not given here. The advantage function is defined as $A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$ [37]. Eq.(4) then becomes:

$$\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a_t \mid s_t) A(s_t, a_t)] \tag{5}$$

This can be viewed as a special case of the actor-critic model, where $\pi_\theta(a_t \mid s_t)$ is the actor and the $A(s_t, a_t)$ is the critic. To reduce the number of required parameters, the parameterized temporal difference (TD) error $\delta_\omega = r_t +$

$\gamma V_\omega(S_{s+1}) - V_\omega(S_s)$ can be used to approximate the advantage function [45]. We use two different symbols θ and ω to denote the actor and critic function, but most of these parameters are shared in a main stream neural network, then separated to two branches for policy and value predictions, respectively.

3.2 Off-policy actor-critic with experience replay

On-policy RL methods update the model with the samples collected via the current policy. The experience replay (ER) can be used to improve the sample-efficiency[54], where the experiences are randomly sampled from a replay pool $\mathcal{P}$. This ensure the training stability by reducing the data correlation. Since these past experiences were collected from different policies, the use of ER leads to off-policy updates.

When training models with RL, ε-greedy action selection is often used to trade-off between exploitation and exploration, whereby a random action is chosen with a probability otherwise the top-ranking action is selected. A policy used to generate a training weight is referred to as a behavior policy μ, in contrast to the policy to-be optimized which is called the target policy π.

The basic advantage actor-critic (A2C) training algorithm described in Sec. 3.1 is on-policy, as it assume the actions are drawn from the same policy as the target to-be optimized (i.e., $\mu = \pi$). However, the current policy π is updated with the samples generated from old behavior policies μ in off-policy learning. Therefore, an importance sampling (IS) ratio is used to rescale each sampled reward to correct the sampling bias at time-step t: $\rho_t = \pi(a_t \mid s_t)/\mu(a_t \mid s_t)$ [55]. For A2C, the off-policy gradient for the parametrized state only value function V_ω thus has the form:

$$\Delta\omega^{\text{off}} = \sum_{t=1}^{T}(\bar{R}_t - \hat{V}_\omega(s_t))\nabla_\omega \hat{V}_\omega(s_t)\prod_{i=1}^{t}\rho_i \tag{6}$$

where $\bar{R}_t$ is the off-policy Monte-Carlo return [56]:

$$\bar{R}_t = r_t + \gamma r_{t+1}\prod_{i=1}^{1}\rho_i + \cdots + \gamma^{T-t} r_T \prod_{i=1}^{T-t}\rho_{t+i} \tag{7}$$

Likewise, the updated gradient for policy π_θ is:

$$\Delta\theta^{\text{off}} = \sum_{t=1}^{T}\rho_t \nabla_\theta \log\pi_\theta(a_t \mid s_t)\hat{\delta}_\omega \tag{8}$$

where $\hat{\delta}_\omega = r_t + \gamma\hat{V}_\omega(s_{t+1} - \hat{V}_\omega(s_t)$ is the TD error using the estimated value of $\hat{V}_\omega$.

Here, we introduce a modified Trust Region Policy Optimization method [46, 47]. In addition to maximizing the cumulative reward $J(\theta)$, the optimization is also subject to a Kullback-Leibler (KL) divergence limit between the updated policy θ and an average policy θ_a to ensure safety. This average policy represents a running average of past policies and constrains the updated policy from deviating too far from the average $\theta_a \leftarrow [\alpha\theta_a + (1-\alpha)\theta]$ with a weight α. Thus, given the off-policy policy gradient $\Delta\theta^{\text{off}}$ in Eq.(8), the modified policy gradient with trust region z is calculated as follows:

$$\begin{aligned} \underset{z}{minimize} \quad & \frac{1}{2}\|\Delta\theta^{\text{off}} - z\|_2^2, \\ \text{Subject to:} & \nabla_\theta D_{KL}[\pi_{\theta_a}(s_t)\|\pi_\theta(s_t)]^{\mathrm{T}}\ z \leq \xi \end{aligned} \tag{9}$$

where π is the policy parametrized by θ or θ_a, and ξ controls the magnitude of the KL constraint. Since the constraint is linear, a closed form solution to this quadratic programming problem can be derived using the KKT conditions. Setting $k = \nabla_\theta D_{KL}[\pi_{\theta_a}(s_t)\|\pi_\theta(s_t)]$, we get:

$$z_{tr}^* = \Delta\theta^{\text{off}} - max\left\{\frac{k^{\mathrm{T}}\Delta\theta^{\text{off}} - \xi}{\|k\|_2^2}, 0\right\}k \tag{10}$$

This direction is also shown to be closely related to the natural gradient [57, 58]. The above enhancements speed up and stabilize our A2C network training.

3.3 Pose-guided inter-set dependency model

To model the inter-set dependency without paired-input, we propose a pose-guided stochastic routing scheme. Such a divide-and-conquer idea originated in [59], which constructs several face detectors to charge each view. Given a set of face image, we extract its general feature aggregation $\mathcal{F}_0$, as well as the aggregation of the frontal face features $\mathcal{F}_1$ and profile face feature $\mathcal{F}_2$. The $\mathcal{F}_1$ and $\mathcal{F}_2$ are the weighted average of the features from the near-frontal face images ($\leq 30°$) and profile face images ($> 30°$) respectively, in which the attention is assigned with the observation of full set. We use PIFA [60] to estimate the yaw angle. The sum of weights of the frontal and profile features p_1 and p_2 are with respect to the quality of each pose group. Considering the mirror transforms in data augmentation and the symmetry property of human faces, we do not discriminate the right face and the left face. With PGR, the distance d between two sets of samples is computed as:

$$d = \frac{1}{2}S(\mathcal{F}_0^1, \mathcal{F}_0^2) + \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} S(\mathcal{F}_i^1, \mathcal{F}_j^2)p_i^1 p_j^2 \tag{11}$$

where S is the L2 distance function to measure the distance between two feature vectors. We treat the generic features and pose-specific features equally, and fuse them for evaluations. The number of distance evaluations is decreased to $\mathcal{O}(5n)$.

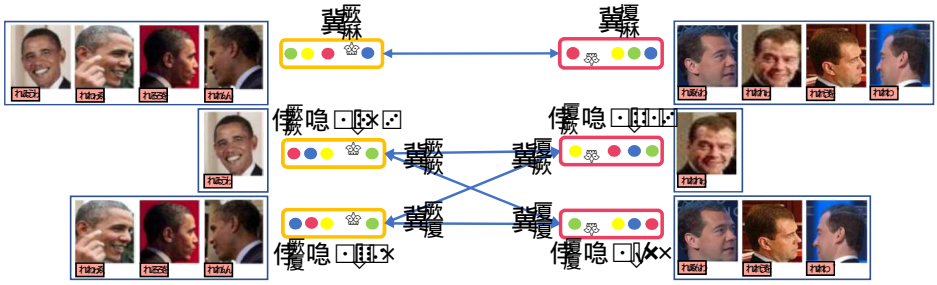


Fig. 4. Illustration of the pose-guided representation scheme.

This achieves promising verification performance requiring fewer comparisons than conventional image-level similarity measurements. It is also readily applied to the other variations.

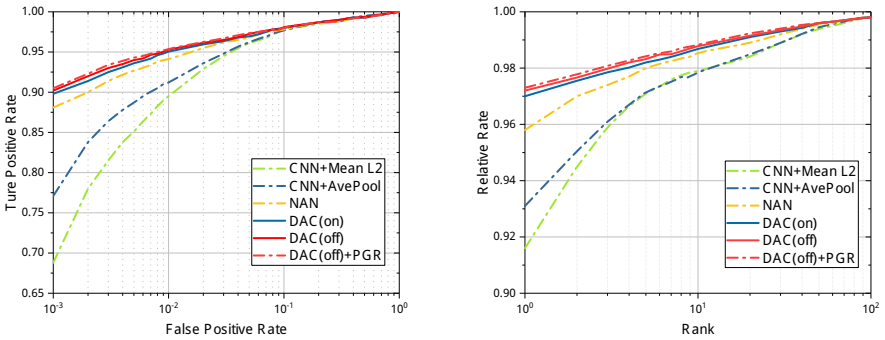
4 Numerical Experiments

We evaluated the performance of the proposed method on three Set/video-based FR datasets: the IJB-A [9], YTF[4], and Celebrity-1000[61]. To utilize the millions of available still images, we train our CNN embedding module separately. As in [16], 3M face images from 50K identities are detected with the JDA [62] and aligned using the LBF [63] method for our GoogleNet training. This part is fixed when we train the DAC module on each set/video face dataset. Benefiting from the highly-compact 128-d feature representation and the simple neural network of the DAC, the training time of our DAC(off) on IJB-A dataset with a single Xeon E5 v4 CPU is about 3 hours, the average testing time per each set-pair for verification is 62ms. We use Titan Xp for CNN processing.

As our baseline methods, CNN+Mean L2 measures the average L2 distances of all image pairs of two sets, while the CNN+AvePool uses average pooling along each feature dimension for aggregation. The previous work NAN [16] uses the same CNN structure as our framework, but adopts a neural network module for independently quality assessment of each image. Therefore, NAN can be also regarded as our baseline. We refer the vanilla A2C as DAC(on), and use DAC(off) for the actor-critic with trust region-based experience replay scheme. The DAC(off)+PGR is the combination of the DAC(off) and PGR.

4.1 Results on IJB-A dataset

IJB-A [9] is a face *verification* and *identification* dataset, containing images captured from unconstrained environments with wide variations of pose and imaging conditions. There are 500 identities with a total of 25,813 images (5,397 still images and 20,412 video frames sampled from 2,042 videos). A set of images for a particular identity is called a template. Each template can be a mixture of



(a) (b)

Fig. 5. Average ROC (Left) and CMC (Right) curves of the proposed method and its baselines on the IJB-A dataset over 10 splits.

still images and sampled video frames. The number of images (or frames) in a template ranges from 1 to 190 with approximately 11.4 images and 4.2 videos per subject on average. It provides a ground truth bounding box for each face with 3 landmarks. There are 10 training and testing splits. Each split contains 333 training and 167 testing identities.

We compare the proposed framework with the existing methods on both face verification and identification following the standard evaluation protocol on IJB-A dataset. Metrics for the 1:1 compare task are evaluated using the receiver operating characteristics (ROC) curves in Fig. 5 (a). We also report the true accept rate (TAR) *vs.* false positive rates (FAR) in Table 1. For the 1:N search task, the performance is evaluated in terms of a Cumulative Match Characteristics (CMC) curve as shown in Fig. 5 (b). It is an information retrieval metric, which plots identification rates corresponding to different ranks. A rank-k identification rate is defined as the percentage of probe searches whose gallery match is returned with in the top-k matches. The true positive identification rate (TPIR) *vs.* false positive identification rate (FPIR) as well as the rank-1 accuracy are also reported in Table 1.

These results show that both the verification and the identification performance are largely improved compared to our baseline methods. The RL networks have learned to be robust to low-quality and redundant image. The DAC(on) outperforms the previous approaches in most of the operating points, showing that our representation is more discriminative than the weighted feature in [16, 11] without considering the inner-set dependency. The experience replay can further help the stabilization of our training and the state-of-the-art performance is achieved. Combining the off-policy DAC and pose-guide representation scheme also contributes to the final results in an efficient way.

Table 1. Performance evaluation on the IJB-A dataset. For verification, the true accept rates (TAR) vs. false positive rates (FAR) are reported. For identification, the true positive identification rate (TPIR) vs. false positive identification rate (FPIR) and the Rank-1 accuracy are presented.

Method	1:1 Verification TAR		1:N Identification TPIR		
	FAR=0.01	FAR=0.1	FPIR=0.01	FPIR=0.1	Rank-1
B-CNN[15]	-	-	0.143±0.027	0.341±0.032	0.588±0.02
LSFS[64]	0.733±0.034	0.895±0.013	0.383±0.063	0.613±0.032	0.820±0.024
DCNN[14]	0.787±0.043	0.947±0.011	-	-	0.852±0.018
Pose-model[65]	0.826±0.018	-	-	-	0.840±0.012
Masi *et al.*[66]	0.886	-	-	-	0.906
Adaptation[6]	0.939±0.013	0.979±0.004	0.774±0.049	0.882±0.016	0.928±0.010
QAN[11]	0.942±0.015	0.980±0.006	-	-	-
NAN[16]	0.941±0.008	0.978±0.003	0.817±0.041	0.917±0.009	0.958±0.005
DAC(on)	0.951±0.014	0.980±0.016	0.852±0.048	0.931±0.012	0.970±0.011
DAC(off)	0.953±0.009	**0.981±0.013**	0.853±0.033	0.933±0.006	0.972±0.012
DAC(off)PGR	**0.954±0.01**	**0.981±0.008**	**0.855±0.042**	**0.934±0.009**	**0.973±0.011**

4.2 Results on YouTube Face dataset

The YouTube Face (YTF) dataset [4] is a widely used video face *verification* dataset, which contains 3,425 videos of 1,595 different subjects. In this dataset, there are many challenging videos, including amateur photography, occlusions, problematic lighting, pose and motion blur. The length of face videos in this dataset varies from 48 to 6,070 frames, and the average length of videos is 181.3 frames. In experiments, we follow the standard verification protocol as in [16, 22, 33], which test our method for unconstrained face 1:1 verification with the given 5,000 video pairs. These pairs are equally divided into 10 splits, and each split has around 250 intra-personal pairs and 250 inter-personal pairs.

Table 2 presents the results of our DAC and previous methods. It can be seen that the DAC outperforms all the previous state-of-the-art methods following the setting that without fine-tuning the feature embedding module on YTF. Since this dataset has frontal face bias [6] and the face variations in this dataset are relatively small as shown in Fig. 2, we have not used the pose-guided representation scheme. It is obvious that the video sequences are redundant, considering the inner-video relationship does contribute to the improvement over [16]. The comparable performance with temporal representation-based methods suggests the DAC could be a potential substitute for RNN in some specific areas. Actually, the RNN itself is computationally expensive and sometimes difficult to train [67]. We directly model the dependency in the feature-level, which is faster than the temporal representation of original images [22], and more effective than the adversarial face generation-based method [33].

It also indicates that DAC achieves a very competitive performance without highly-engineered CNN models. Note that the FaceNet [18], NAN [16] also use the GoogleNet style structure. We show that DAC outperforms them on

Table 2. Comparisons of the average verification accuracy with the recently state-of-the-art results on the YTF dataset.† fine-tuned the CNN model with YTF.

Method	Accuracy	†Accuracy	Year
FaceNet[18]	0.9512±0.0039	-	2015
Deep FR[13]	0.915	0.973	2015
CenterLoss[20]	0.949	-	2016
TBE-CNN[68]	0.9384±0.0032	0.9496±0.0031	2017
TR[22]	0.9596±0.0059	0.9652±0.0054	2017
NAN[16]	0.9572±0.0064	-	2017
DAN[33]	0.9428±0.0069	-	2017
DAC(on)	0.9597±0.0041		
DAC(off)	**0.9601±0.0048**		

both the verification accuracy and the standard variation. The Deep FR, TBE-CNN and TR methods have additional fine-tuning of the CNN-model with YTF dataset, and the residual constitutional networks are used in TR. Considering our module-based structure, these advanced CNNs can be easily added on the DAC and boost its performance. We see that the DAC can generalizes well in video-based face verification datasets.

4.3 Results on Celebrity-1000 dataset

We then test our method on the Celebrity-1000 dataset [61], which is designed for the unconstrained video-based face *identification* problem. 2.4M frames from 159,726 face videos (about 15 frames per sequence) of 1,000 subjects are contained in this dataset. It is released with two standard evaluation protocols: open-set and closed-set. We follow the standard $1 : N$ identification setting as in [61] and report the result of both protocols.

For the closed-set protocol, we use the softmax outputs from the reward network, and the subject with the maximum score as the result. Since the baseline methods do not have a multi-class prediction unit, we simply compare the L2 distance as in [16]. We present the results in Table 3, and show the CMC curves in Fig. 6 (a). With the help of end-to-end learning and large volume training data for CNN model, deep learning methods outperform [61, 12] by a large margin. It can be seen that the state-of-the-art is achieved by the DAC. We can also benefit from the experience replay to achieve improvements over the baselines.

For the open-set testing, we take multiple image sequences of each gallery subject to extract a highly compact feature representation as in NAN [16]. Then the open-set identification is performed by comparing the L2 distance of the aggregated probe and gallery representations. Fig. 6 (b) and Table 3 show the results of different methods in our experiments. We see that our proposed methods outperform the previous methods again, which clearly shows that DAC is effective and robust.

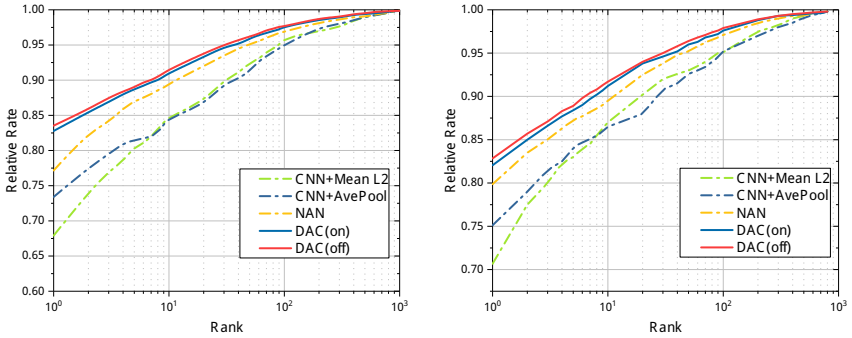


Fig. 6. The CMC curves of different methods on Celebrity 1000. (a) Close-set tests on 1000 subjects, (b) Open-set tests on 800 subjects

.

Table 3. Identification performance (rank-1 accuracies), on the Celebrity-1000 dataset for the closed-set tests (left) and open-set tests (right).

Method	Number of Subjects(*closed*)				Number of subjects(*open*)			
	100	200	500	1000	100	200	500	800
MTJSR[61]	0.506	0.408	0.3546	0.3004	0.4612	0.3984	0.3751	0.3350
Eigen-PEP[12]	0.506	0.4502	0.3997	0.3194	0.5155	0.4615	0.4233	0.2590
CNN+Mean L2	0.8526	0.7759	0.7457	0.6791	0.8488	0.7988	0.7676	0.7067
CNN+AvePool	0.8446	0.7893	0.7768	0.7341	0.8411	0.7909	0.7840	0.7512
NAN[16]	0.9044	0.8333	0.8227	0.7717	0.8876	0.8521	0.8274	0.7987
DAC(on)	0.9125	0.8722	0.8475	0.8278	0.8986	0.8706	0.8395	0.8205
DAC(off)	**0.9137**	**0.8783**	**0.8523**	**0.8353**	**0.9004**	**0.8715**	**0.8428**	**0.8264**

5 Conclusions

We have introduced the actor-critic RL for visual recognition problem. We cast the inner-set dependency modeling to a MDP, and train an agent DAC to make attention control for each image in each step. The PGR scheme well balances the computation cost and information utilization. Although we only explore their ability in set/video-based face recognition tasks, we believe it is a general and practicable methodology that could be easily applied to other problems, such as Re-ID, action recognition and event detection *etc*.

6 Acknowledgement

This work was supported in part by the National Key R&D Plan 2016YFB0501003, Hong Kong Government General Research Fund GRF 152202/14E, PolyU Central Research Grant G-YBJW, Youth Innovation Promotion Association, CAS (2017264), Innovative Foundation of CIOMP, CAS (Y586320150).

References

1. Chen, J.C., Ranjan, R., Sankaranarayanan, S., Kumar, A., Chen, C.H., Patel, V.M., Castillo, C.D., Chellappa, R.: Unconstrained still/video-based face verification with deep convolutional neural networks. International Journal of Computer Vision (2017) 1–20
2. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: A survey. In: Advances in face detection and facial image analysis. Springer (2016) 189–248
3. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst (2007)
4. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. (2011) 529–534
5. Phillips, P.J., Hill, M.Q., Swindle, J.A., O'Toole, A.J.: Human and algorithm performance on the pasc face recognition challenge. In: Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on, IEEE (2015) 1–8
6. Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., Zisserman, A.: Template adaptation for face verification and identification. In: FG, IEEE (2017) 1–8
7. Hayat, M., Khan, S.H., Werghi, N., Goecke, R.: Joint registration and representation learning for unconstrained face identification. In: IEEE CVPR. (2017) 2767–2776
8. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: IEEE CVPR. Volume 1. (2017)
9. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1931–1939
10. Grother, P., Ngan, M.: Face recognition vendor test (frvt). Performance of face identification algorithms (2014)
11. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit. (2017) 5790–5799
12. Li, H., Hua, G., Shen, X., Lin, Z., Brandt, J.: Eigen-pep for video face recognition. In: Asian Conference on Computer Vision, Springer (2014) 17–33
13. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC. Volume 1. (2015) 6
14. Chen, J.C., Ranjan, R., Kumar, A., Chen, C.H., Patel, V.M., Chellappa, R.: An end-to-end system for unconstrained face verification with deep convolutional neural networks. In: IEEE CVPRW. (2015) 118–126
15. Chowdhury, A.R., Lin, T.Y., Maji, S., Learned-Miller, E.: One-to-many face recognition with bilinear cnns. In: WACV, IEEE (2016) 1–9
16. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: IEEE CVPR. (2017) 4362–4371
17. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 1701–1708
18. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 815–823

19. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 2892–2900
20. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision, Springer (2016) 499–515
21. Zhang, J., Wang, N., Zhang, L.: Multi-shot pedestrian re-identification via sequential decision making. arXiv preprint arXiv:1712.07257 (2017)
22. Rao, Y., Lu, J., Zhou, J.: Attention-aware deep reinforcement learning for video face recognition. In: IEEE ICCV. (2017) 3931–3940
23. Janisch, J., Pevnỳ, T., Lisỳ, V.: Classification with costly features using deep reinforcement learning. arXiv preprint arXiv:1711.07364 (2017)
24. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: a deep model for learning face identity and view representations. In: Advances in Neural Information Processing Systems. (2014) 217–225
25. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence **31**(2) (2009) 210–227
26. Cevikalp, H., Triggs, B.: Face recognition based on image sets. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2567–2573
27. Huang, Z., Wu, J., Van Gool, L.: Building deep networks on grassmann manifolds. arXiv preprint arXiv:1611.05742 (2016)
28. Huang, Z., Van Gool, L.J.: A riemannian network for spd matrix learning. In: AAAI. Volume 2. (2017) 6
29. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2496–2503
30. Lu, J., Wang, G., Moulin, P.: Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 329–336
31. Sivic, J., Everingham, M., Zisserman, A.: who are you?-learning person specific classifiers from video. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1145–1152
32. Lu, J., Wang, G., Deng, W., Moulin, P., Zhou, J.: Multi-manifold deep metric learning for image set classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1137–1145
33. Rao, Y., Lin, J., Lu, J., Zhou, J.: Learning discriminative aggregation network for video-based face recognition. In: IEEE ICCV. (2017) 3781–3790
34. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
35. Vinyals, O., Bengio, S., Kudlur, M.: Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391 (2015)
36. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540) (2015) 529
37. Li, Y.: Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274 (2017)
38. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. In: ICML. (2014)

39. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)
40. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning. (2016) 1928–1937
41. Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., Kautz, J.: Reinforcement learning through asynchronous advantage actor-critic on a gpu. (2017)
42. Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.: Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine **34**(6) (2017) 26–38
43. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: NIPS. (2000) 1057–1063
44. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. In: Reinforcement Learning. Springer (1992) 5–32
45. Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation. (2017)
46. Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P.: Trust region policy optimization. In: ICML. (2015) 1889–1897
47. Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., de Freitas, N.: Sample efficient actor-critic with experience replay. (2017)
48. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in neural information processing systems. (2014) 2204–2212
49. Lan, X., Wang, H., Gong, S., Zhu, X.: Identity alignment by noisy pixel removal. arXiv preprint arXiv:1707.02785 (2017)
50. Huang, C., Lucey, S., Ramanan, D.: Learning policies for adaptive tracking with deep feature cascades. arXiv preprint arXiv:1708.02973 (2017)
51. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2818–2826
52. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. (2015) 448–456
53. Andrew, A.M.: Reinforcement learning: An introduction by richard s. sutton and andrew g. barto, adaptive computation and machine learning series, mit press (bradford book), cambridge, mass., 1998, xviii+ 322 pp, isbn 0-262-19398-1,(hardback,£ 31.95).-. Robotica **17**(2) (1999) 229–235
54. Lin, L.J.: Self-improving reactive agents based on reinforcement learning, planning and teaching. Machine learning **8**(3-4) (1992) 293–321
55. Meuleau, N., Peshkin, L., Kaelbling, L.P., Kim, K.E.: Off-policy policy search. MIT Articical Intelligence Laboratory (2000)
56. Precup, D., Sutton, R.S., Dasgupta, S.: Off-policy temporal-difference learning with function approximation. In: ICML. (2001) 417–424
57. Amari, S.I.: Natural gradient works efficiently in learning. Neural computation **10**(2) (1998) 251–276
58. Peters, J., Schaal, S.: Policy gradient methods for robotics. In: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on, IEEE (2006) 2219–2225
59. Li, Y., Zhang, B., Shan, S., Chen, X., Gao, W.: Bagging based efficient kernel fisher discriminant analysis for face recognition. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Volume 3., IEEE (2006) 523–526

60. Jourabloo, A., Liu, X.: Pose-invariant face alignment via cnn-based dense 3d model fitting. International Journal of Computer Vision **124**(2) (2017) 187–203
61. Liu, L., Zhang, L., Liu, H., Yan, S.: Toward large-population face identification in unconstrained videos. IEEE Transactions on Circuits and Systems for Video Technology **24**(11) (2014) 1874–1884
62. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: European Conference on Computer Vision, Springer (2014) 109–122
63. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1685–1692
64. Wang, D., Otto, C., Jain, A.K.: Face search at scale: 80 million gallery. arXiv preprint arXiv:1507.07242 (2015)
65. Masi, I., Rawls, S., Medioni, G., Natarajan, P.: Pose-aware face recognition in the wild. In: IEEE CVPR. (2016) 4838–4846
66. Masi, I., Trn, A.T., Hassner, T., Leksut, J.T., Medioni, G.: Do we really need to collect millions of faces for effective face recognition? In: ECCV, Springer (2016) 579–596
67. Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C.L.Y., Courville, A.: Towards end-to-end speech recognition with deep convolutional neural networks. arXiv preprint arXiv:1701.02720 (2017)
68. Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE transactions on pattern analysis and machine intelligence (2017)