

Discriminative Invariant Kernel Features: A Bells-and-Whistles-Free Approach to Unsupervised Face Recognition and Pose Estimation

Dipan K. Pal
 Carnegie Mellon University
 dipanp@andrew.cmu.edu

Felix Juefei-Xu
 Carnegie Mellon University
 felixu@cmu.edu

Marios Savvides
 Carnegie Mellon University
 msavvid@ri.cmu.edu

Abstract

We propose an explicitly discriminative and ‘simple’ approach to generate invariance to nuisance transformations modeled as unitary. In practice, the approach works well to handle non-unitary transformations as well. Our theoretical results extend the reach of a recent theory of invariance to discriminative and kernelized features based on unitary kernels. As a special case, a **single** common framework can be used to generate subject-specific pose-invariant features for face recognition and vice-versa for pose estimation. We show that our main proposed method (DIKF) can perform well under very challenging large-scale semi-synthetic face matching and pose estimation protocols with unaligned faces using no landmarking whatsoever. We additionally benchmark on CMU MPIE and outperform previous work in almost all cases on off-angle face matching while we are on par with the previous state-of-the-art on the LFW unsupervised and image-restricted protocols, **without** any low-level image descriptors other than **raw-pixels**.

1. Introduction

Over the years there have been many different approaches to the problem of general face recognition and pose estimation. However, there is a lot yet to be achieved before one can consider the unconstrained version of these problems fully solved. Though there are many algorithms in literature that perform extremely well on unconstrained datasets such as the LFW [21, 35, 8, 40, 6, 26, 38], given the complexity of these algorithms, it remains unclear as to what underlying objective each of them aims to achieve in the context of unconstrained face matching. To address this need, we refrain from incorporating ‘complex’ algorithms and optimization problems in our approach and solely base the core mechanism on fundamental principles of generating invariance.

Although success on the LFW framework has been very encouraging, a paradigm shift is underway towards the role

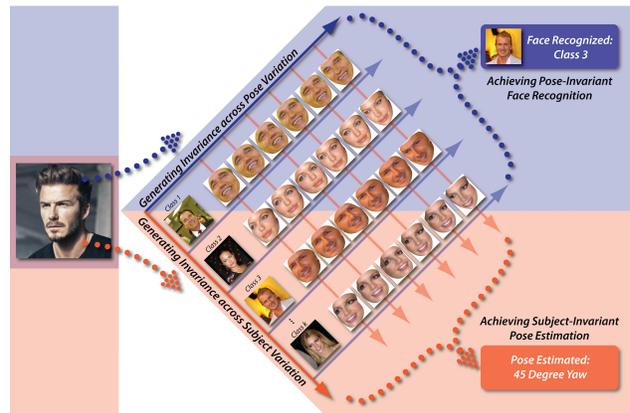


Figure 1: The common single framework for pose-invariant face recognition and subject-invariant pose estimation. Measuring moments (pooling) across a transformation (*i.e.* subject or pose in this figure) invokes invariance towards that transformation.

of such large unconstrained databases. It has been suggested, that the problem of face recognition be divided into subtasks of achieving invariance towards transformations of a face [23]. Further, strong recent progress on the supervised protocols of the LFW dataset nearing human accuracy have perhaps overshadowed the need to understand the fundamental problems in vision tasks such as recognition.

In light of these arguments, there is a need for methods that are based on fundamental principles, not in order to beat the current state-of-the-art, but instead to further our understanding of the problem itself. Although, there has been significant work in literature generating implicit invariance to specific individual or a small subset of these transformations at once, an approach which generates *explicitly invariant* features to any unitary modeled transformation while being *explicitly discriminative* has not been studied. Further, there is a need for a study investigating generation of invariant features to multiple common transformations of faces in a controlled setting. Studies on large-scale controlled datasets have the advantage of being more informative regarding the algorithmic shortcomings.

Contributions. In this paper, we present an approach to obtain explicitly discriminative features that are invariant to multiple transformations that can be locally modeled as unitary. The approach does not involve solving a heavy optimization problem. It is based, instead, on a different manifestation of group invariance and on learning discriminative filters through a simple closed form solution. We focus part of our study to be more exploratory towards understanding the challenges due to specific common transformations such as pose, translation, scale and rotation. Even using a simple method with a closed-form solution, working only with *raw-pixels*, we are able to, to the best of our knowledge, achieve the state-of-the-art on the MPIE database protocol, and match the previous state-of-the-art on the LFW unsupervised and image-restricted label-free protocol. Hence our approach is ‘bells and whistles free’. Our main contributions are:

1. We propose a simple approach to learn *discriminative* non-linear features that are invariant to unitary transformations. We extend the reach of a recent theory of invariance to discriminative and kernelized features.
2. Focusing on the challenging transformation of pose, we propose a *simple* dense-landmark-free approach which results in a framework capable of performing *open-set* pose-invariant face recognition and *simultaneous* pose estimation as illustrated in Fig. 1.
3. We extend the method to result in a sequential approach of generating invariance to multiple sub-groups of transformations. Using this, we arrive at a completely landmark-free framework (at evaluation) for transformation invariant face recognition and pose estimation.

Related work. A majority of the recent efforts on unconstrained face verification focus on LFW, and rely greatly on locating accurate and dense facial landmarks and descriptors to extract overcomplete information from the image and/or use 3D modeling in the algorithm [35, 8, 25, 40, 7, 6, 31, 11, 3, 24, 42, 22, 38]. Many of these systems are also closed set, whereas our approach is inherently *open-set*. Further they address different types of transformations differently. Automatic dense facial landmarking, in fact already factors out a majority of the transformations such as translation, rotation and scale. Further, LFW provides reasonably aligned images which help to factor out translation, in-plane rotation and scale. This enables the algorithms to escape the need to account for those transformations within the core framework. In practice, real-time landmarking is expensive, thus there arises a need to explore methods than circumvent the requirement for dense and accurate landmarks.

A slightly different class of algorithms based on deep learning have gained popularity recently, which utilizes a lot of data (high sample complexity) and increases model complexity drastically [37, 36, 32, 12]. These methods although widely successful, fail to provide a better understanding of the problem due to complex models and over-complete feature extraction combined with unconstrained testing protocols.

In light of the current trend in unconstrained face recognition, large-scale databases such as LFW include an uncontrolled amount of certain unspecified types of transformations in each image. However, other transformations such as translation, in-plane rotation and scale are factored out by providing aligned faces. Having no control over the type and amount of other transformations tends to bias the development of face recognition systems where it is not clear why some algorithms work well while others don’t. Thus, eluding the underlying problem which is to generate *invariance* to all intra-class transformations while being *discriminative* amongst inter-class transformations. In this work, we present results on a semi-synthetic large-scale data with controlled amounts of transformations.

Normalized invariant dot-products (NDP). Some recent work by Liao *et al.* [26], adopting a perspective similar to ours in this paper, perform competitively on LFW. However, they use ‘external training data’ to extract features and require accurate facial alignment before extracting more complex feature descriptors (HOG followed by PCA). To have a fair comparison in our exploratory experiments, we compare against their core baseline method of using normalized dot-products followed by mean (NDP- l_1) and max (NDP- l_∞) pooling. The multilayer extension of our approach on the other hand requires no such alignment for recognition of misaligned faces and we restrict ourselves to work with raw pixels throughout this study.

2. Linear Invariant Random Features

I-theory [2] was proposed to generate invariance to transformations motivated by the properties of the visual cortex. Empirically, through experiments on LFW, it has been shown that sufficient invariance can be learned [26].

Sample complexity. One of the main motivations of I-theory of invariance is the problem of reducing sample complexity to learn a concept. It can be argued that humans tend to learn new concepts (*e.g.* the structure of a novel object) with very few examples. Yet most machine learning and vision algorithms require a lot of data to learn, with the general focus being on performance rather than sample complexity. Given the advances in cheaper computation, sample complexity might not seem important. However, in order to understand the low sample complexity characteristics of human vision, it might be useful to better explore paradigms which try to achieve low sample complexity. I-

theory is one such paradigm we now briefly overview.

Consider a unitary group of transformations G with group elements g with finite cardinality ($|G|$). One can represent the action (*i.e.* translation, rotation *etc.*) of the group element g on an image as $gI(x) = I(g^{-1}x)$. The *orbit* of the images I' generated by $\{I' \mid I' = g(I) \forall g \in G\}$ is *unique* to every image since it is the set of all variations of an image as defined by G . In order to compare two orbits, a measure which also introduces invariance is the probability distribution P_I induced by the group elements of G on I . The heuristic is to directly/indirectly measure a statistic of the probability distribution, thus generating invariance. It can be shown that $I \sim I' \Leftrightarrow P_I = P_{I'}$, *i.e.* if two images are equivalent under some g , then their distributions are identical [2]. To characterize the distribution P_I , one can use *arbitrary* templates $\mathbf{t}^k, k \in \{1, \dots, K\}$, each template providing a 1-D projection of P_I . In order to achieve discriminability between say n images/orbits, up to precision ϵ , with confidence $(1 - \delta)$, one must have $K \geq \frac{2}{\epsilon\delta} \log \frac{n}{\delta}$ [2]. Perhaps one of the most interesting observations that I-theory makes is that, since g is unitary, we have:

$$\langle g(I), \mathbf{t} \rangle = \langle I, g^{-1}(\mathbf{t}) \rangle \quad (1)$$

Thus, the distribution of the set $\{\langle gI, \mathbf{t} \rangle\}, \forall g \in G$ is the same as that of $\{\langle I, g^{-1}\mathbf{t} \rangle\}, \forall g \in G$. Hence, it is not necessary to explicitly observe all transformations of a novel image in order to discriminate it. This is the key to reduce sample complexity, and what our approach capitalizes on working with few faces (~ 3000 for LFW and <350 for other experiments).

One can then use multiple 1-D projections of the distribution in order to characterize the novel object. Characterizing this distribution leads to an invariant computed through a histogram $\mu^n(I) = \frac{1}{|K|} \sum_k \eta(\langle I, g_k(\mathbf{t}_n) \rangle), \forall n = 1, \dots, N$. Here, η can be a non-linear threshold function $\eta : \mathbb{R} \mapsto \mathbb{R}$ which results in a histogram approximation of the distribution. The distribution of $\{\langle gI, \mathbf{t} \rangle\}, \forall g \in G$ can also be explained by its set of moments. Liao *et al.* [26] found that the first moment approximation of the distribution worked well in practice which translates to mean pooling. Different pooling schemes would capture different aspects (moments) of the 1-D distribution.

3. Discriminative Invariant Kernel Features

Discriminative templates to generate invariant features. Section 2 describes how a set of arbitrary templates transformed by an unitary group can be used to invoke invariance to that group through statistics of the invariant distribution under the action of the group. However, the templates \mathbf{t} are not discriminative and they may capture redundant information about the distribution. Discriminability

between classes may offer better separation between the orbits of different images.

Discriminative invariant linear features (DILF). We adopt a simple way of generating discrimination between K classes. We seek a filter or template \mathbf{t}_k s.t. $\mathbf{X}^T \mathbf{t}_k = \mathbf{u}_k$, where $\mathbf{X} \in \mathbb{R}^{d \times K}$ is the pre-whitened data matrix with K classes. \mathbf{u}_k is a label vector of zeros with 1 at position k for the k -th class. Under the constraint that the template has to be a linear combination of the data, the solution becomes $\mathbf{t}_k = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{u}_k$. Now, consider a finite unitary group G acting on \mathbf{X} , thereby generating $\{g_n(\mathbf{X}_n)\}$ *i.e.* $\mathbf{X}_n = g_n(\mathbf{X})$, where every element in \mathbf{X} has been acted upon by $g_n \in G$ ¹. We train $N \times K$ separate templates or filters \mathbf{t}_{kn} , one for each class and each transformation. Concretely, $\mathbf{t}_{kn} = \mathbf{X}_n(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{u}_k$.

A key issue is that the learned templates need to be the action of a unitary group in order for the distribution under the orbit to be invariant to that group. Specifically, if $\mathbf{T} = \{\mathbf{t}_k\}$ is the set of learned templates or filters then, it should be possible to express \mathbf{T} as $\{g_n(\mathbf{t}_0) \mid g_n \in G\}$. We find that this is indeed the case.

Theorem 3.1 (DILF filters form a set of transformed templates under a group). *Given a group G of unitary transformation elements g with $|G| = N$ and $\{\mathbf{X}_n \mid n = 1, \dots, N\}$ are pre-whitened template matrices, then the set of DILF filters $\mathcal{T}_k = \{\mathbf{t}_{kn} = \mathbf{X}_n(\mathbf{X}_n^* \mathbf{X}_n)^{-1} \mathbf{u}_k \mid n = 1, \dots, N\}$ is a set of transformed templates under the action of group G .*

The proof is analogous to that of Theorem 3.2 to be presented later (with a linear kernel). Since all K sets \mathcal{T}_k can be expressed as being acted on by a group, one can compute an invariant feature of dimension K by either estimating the distribution or by computing moments of the K 1-D distributions $\{\langle I, \mathbf{t}_{kn} \rangle\}, \forall k$ [2]. Hence, the feature extraction can capitalize on storing templates to ‘learn’ transformations in group G while being explicitly discriminative.

Although DILF does achieve explicit discriminability, performance can be enhanced by dot-products in a high-dimensional space incorporating inherent non-linearities. This directly motivates the use of kernels into DILF thus arriving at Discriminative Invariant Kernel Features (DIKF).

Discriminative invariant kernel features (DIKF). We now present our central method of extracting explicitly discriminative invariant features. Our goal remains the same as in the previous section, and so does our problem formulation. However, now our approach incorporates high-dimensional embeddings in the form of kernels. We emphasize the fact that the method remains inherently simple (dot-products followed by statistics computed over the invariant distribution) albeit in a much high dimensional space.

¹This is a slight abuse of notation, wherein g can act both on vectors and column-wise on matrices.

Consider a feature mapping $\Phi : \mathbb{R}^d \rightarrow \mathbb{H}$, to some high-dimensional Hilbert space \mathbb{H} , then learning the filter on a data matrix \mathbf{X}_n (where each column vector is a vectorized image $\mathbf{x}_{kn}, \forall k = 1, \dots, K$ with transformation $g_n \in G$), the template becomes

$$\Phi(\mathbf{t}_{kn}) = \Phi(\mathbf{X}_n)(\Phi(\mathbf{X}_n) \cdot \Phi(\mathbf{X}_n))^{-1} \mathbf{u}_k \quad (2)$$

However, in order to be able to extract invariant features using DIKF, the K sets of filters need to form to a set acted upon by group G . This puts constraints on the kind of kernel Φ can be. In this case, Φ is constrained to be a unitary kernel as per the following definition.

Definition 3.1 (Unitary Kernel). We define a kernel $k(x, y) = \langle \phi(x), \phi(y) \rangle$ to be a unitary kernel if, for a unitary group G , the mapping $\phi(x) : \mathcal{X} \rightarrow \mathbb{H}$ satisfies $\langle \phi(gx), \phi(gy) \rangle = \langle \phi(x), \phi(y) \rangle \forall g \in G, \forall x, y \in \mathcal{X}$.

We now show that for unitary kernels, indeed, DIKF filters form a set of transformed templates in the kernel space under the unitary group.

Theorem 3.2 (DIKF filters form a set of transformed templates in the kernel space under a group). *Given a group G of unitary transformation elements g with $|G| = N$, if $k(x, y) = \langle \phi(x), \phi(y) \rangle$ i.e. k is a unitary kernel, and $\{\mathbf{X}_n \mid \mathbf{X}_n = g_n(\mathbf{X}), g_n \in G\}$ are a set of pre-whitened matrices acted upon by G , then the set of DIKF filters*

$$\mathcal{T}_k = \left\{ \Phi(\mathbf{t}_{kn}) = \Phi(\mathbf{X}_n)(\Phi(\mathbf{X}_n) \cdot \Phi(\mathbf{X}_n))^{-1} \mathbf{u}_k \mid \forall n \right\}$$

is a set of transformed templates under a group.

Proof. Without loss of generality, consider $\mathbf{X}_1 = \{\mathbf{x}_i \mid i = 1, \dots, K\}$ and $\mathbf{X}_n = \{g_n(\mathbf{x}_i) \mid i = 1, \dots, K\}$ where $g_n \in G$. Thus, $\Phi(\mathbf{t}_{kn}) = \Phi(\mathbf{X}_n)(\Phi(\mathbf{X}_n) \cdot \Phi(\mathbf{X}_n))^{-1} \mathbf{u}_k$.

For an unitary kernel, $\langle \Phi(g_n(\mathbf{x})), \Phi(g_n(\mathbf{y})) \rangle = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$. Thus, $\exists \bar{g} : \mathbb{H} \rightarrow \mathbb{H}$ s.t. $\langle \bar{g}\Phi(\mathbf{x}), \bar{g}\Phi(\mathbf{y}) \rangle = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$. \bar{g} is the transformation between \mathbf{x} and $g(\mathbf{x})$ in the kernel Hilbert space. The unitary kernel property implies that \bar{g} is unitary and therefore linear. Further, it forms a group G' in the kernel Hilbert space. Now, we have

$$\Phi(\mathbf{t}_{kn}) = \Phi(\mathbf{X}_n)(\Phi(\mathbf{X}_n) \cdot \Phi(\mathbf{X}_n))^{-1} \mathbf{u}_k \quad (3)$$

$$= \Phi(g_n(\mathbf{X}_1))(\Phi(g_n(\mathbf{X}_1)) \cdot \Phi(g_n(\mathbf{X}_1)))^{-1} \mathbf{u}_k \quad (4)$$

$$= \Phi(g_n(\mathbf{X}_1))(\Phi(\mathbf{X}_1) \cdot \Phi(\mathbf{X}_1))^{-1} \mathbf{u}_k \quad (5)$$

$$= \Phi(g_n(\mathbf{X}_1)) \mathbf{v}_k = \bar{g}_n(\Phi(\mathbf{X}_1)) \mathbf{v}_k \quad (6)$$

$$= \bar{g}_n(\Phi(\mathbf{X}_1)) \mathbf{v}_k \quad (7)$$

In Eqn. 4, with a slight abuse of notation, we use the same notation $g_n(\cdot)$ and $\bar{g}_n(\cdot)$ as an operator both on a single image \mathbf{x} or a matrix of images \mathbf{X} . In Eqn. 5, note that the transformation in the kernel Hilbert space is unitary. In Eqn. 6,

we put $\mathbf{v}_k = (\Phi(\mathbf{X}_1) \cdot \Phi(\mathbf{X}_1))^{-1} \mathbf{u}_k$ and recall that \bar{g}_n is linear. Thus, every element of the template set

$$\Phi(\mathbf{t}_{kn}) = \Phi(\mathbf{X}_n)(\Phi(\mathbf{X}_n) \cdot \Phi(\mathbf{X}_n))^{-1} \mathbf{u}_k \in \mathcal{H}_k$$

can be written as an unitary transformation of the vector $\Phi(\mathbf{X}_1) \mathbf{v}_k$ with group element $\bar{g}_n \in G'$. \square

Preservation of the group transformation property for all sets \mathcal{T}_k even in the kernel Hilbert space allows for 1-D distributions of the filter responses \mathbf{t}_{kn} with a novel image to be invariant to G . One can compute statistics, such as moments, that therefore become invariant to G in the original image space. In this paper, we explore two such moments, the first moment translating to mean pooling and the infinity moment translating to max pooling. Hence, we can use the learned filters to model the transformations of the data instead of needing to explicitly observe transformations of the novel image, thereby reducing sample complexity. It is interesting to note that unitary transformations allow sets of non-linear filters (templates or hyperplanes) to form sets of transformed templates under a group in the kernel Hilbert space as well. This would allow a much broader class of discriminative models to fit into the approach of generating invariance through moment measurement. Nonetheless, in this study, we restrict ourselves to filter based approaches.

4. Common Framework for Landmark-Free Unsupervised Pose-Invariant Face Recognition and Pose Estimation

Applying DIKF to faces. Although discriminative invariant feature extraction framework can be applied to any kind of data observing transformations modeled by a group, in this work, we focus on faces. Challenging transformations of faces include translation, in-plane rotation and scaling which can be perfectly modeled linearly by some unitary G . However, out-of-plane rotation or pose variation is considered to be much more challenging being non-linear. Nonetheless, a small enough pose variation can indeed be approximated by some G . It has been previously observed that pose variations can be piece-wise linearly approximated through transformation-dependent submanifold unfolding [29]. The training set $\mathcal{X} = \{\mathbf{X}_n \mid g_n \in G\}$ for training our templates to generate invariance towards pose variation would involve faces in different poses. In the case of faces, we do this by generating a 3-D model of each face template and then rendering them in different poses using 3D generic elastic model (3DGEM) [14]. Thus, for K different faces, we can have N different transformations of pose. The training set would thus have subject variation along one axis and pose variation along the other (see Fig. 1). Note that this step is part of dataset generation and not actually a part of the algorithm.

Unsupervised training of templates. It is straightforward to apply DIKF to a supervised setting. For the unsupervised setting, we simply choose random faces (not restricted to be from different subjects), and generate multiple poses to obtain the training set used to learn the templates/filters. In training DILF and DIKF, we simply allow \mathbf{u}_k to be 1 only for the k -th face, thereby extracting features discriminative between the faces in the training set *without* using any labels.

Single framework for pose-invariant face recognition and pose estimation. There are two different kinds of transformations modeled in the face training set $\mathcal{X} = \{\mathbf{X}_n \mid g_n \in G\}$, *i.e.* pose variation and subject variation. The transformation across subjects is a much harder transformation to model. Even though there exists in general, an affine transformation between subjects, it is hard to prove it is a group, and thus is an abuse of the theoretical framework. Nonetheless, in practice, we explore the limits to the method. To generate invariance, we ‘pool’ across subjects for a subject-invariant pose-selective feature for pose estimation and ‘pool’ across poses to obtain a pose-invariant subject-selective feature for face verification. This is the essence of the common framework. Note that even if we are unable to observe the entire orbit of G , approximate invariance would hold [2]. If we only pool across pose variation, we only require the two eye-center locations to be aligned across all the faces. This is a much less restrictive and computationally feasible condition than having to estimate a dense set of landmarks in order to perform feature extraction [31, 6, 7, 40, 8, 35]. We later drop this requirement and move towards a completely landmark-free system at evaluation. Algorithm 1 formally describes DIKF, where G in our case models pose-variation for face verification.

Algorithm 1: Extracting DIKF for I invariant to G

input : Input image I (vectorized), $\mathcal{X} = \{\mathbf{X}_n = g_n(X) \mid n = 1, \dots, N\}$, $\{\mathbf{u}_k \mid k = 1, \dots, K\}$, Φ , G with $|G| = N$

output: Invariant feature vector $\mu \in \mathbb{R}^K$

- 1 *Learn and compute correlations with filters;*
- 2 **for** $g_n \in G$ **do**
- 3 **for** $k = 1, \dots, K$ **do**
- 4
$$\begin{cases} \mathbf{t}_{kn} = \mathbf{X}_n(\Phi(\mathbf{X}_n) \cdot \Phi(\mathbf{X}_n))^{-1} \mathbf{u}_k \\ f_{kn} = \langle \Phi(\mathbf{t}_{kn})^\top, \Phi(I) \rangle \end{cases}$$
- 5 *Compute first / infinity moment;*
- 6 **for** $k = 1, \dots, K$ **do**
- 7 $\mu_k = \frac{1}{N} \sum_i f_{ki}$ (*Mean Pooling*)
- 8 $\mu_k = \|\mathbf{f}_k\|_\infty$ (*Max Pooling*)

Sequential invariance to multiple transforms. Affine transformations in images span a huge space, and to generate invariance to the entire group would involve dense sam-

pling of the orbit of the group (which comprises of sub-groups *e.g.* unitary transforms contain translation and rotation). This would result in sample complexity being exponential in the number of sub-group elements since we are required to sample all combinations. Instead, we consider the factored-out transformations in *sequence*, *e.g.* the translation sub-group, followed by the rotation sub-group and so on. This reduces the sample complexity drastically down to being linear in number of sub-group elements. The procedure applies Algorithm 1 in sequence in multiple ‘levels’, with each level having its training template based off features from the previous level. Thus, each level progressively adds in invariance until all sub-groups being modeled are covered.

5. Experimental Results

5.1. Pose-Invariant Face Recognition

We test our proposed DIKF for pose-invariant face recognition and compare against NDP in a recent study [26]².

A. Single-level pose-invariant face recognition on a large-scale semi-synthetic mugshot database.

Our first experiment is exploratory, and deals exclusively with the most important transformation considered in this paper, *i.e.*, pose variation. Although there has been some interesting previous work on testing pose invariance exclusively using NDP, the experiments were carried out in small scale and by using fully synthetic faces with minimal texture variation [2]. Even though strong results using a similar pooling approach on LFW have been recently established, the existence of a multitude of transformations in LFW make it difficult to establish how effective was the invariance to pose transformation specifically [26].

Dataset generation. We choose to focus exclusively on the transformation of *pose*, thus we design this experiment to *not* include most other transformations such as translation, scale and in-plane rotation. We start with 1,000 frontal mugshot images of different subjects and then use 3D-GEM to generate a 3D model of each face and render multiple poses. We render poses varying from -40° to 40° (yaw) and -20° to 20° (pitch) in steps of 5° . Thus, in all for each subject we obtain 153 different poses, thereby coming up to a total of 153,000 images³.

Protocol. We evaluate ℓ_∞ -DIKF and ℓ_1 -DIKF and

²We present more experiments in the supplementary material.

³The dataset that we generate can be termed as a semi-synthetic dataset, since we use real-world face images to render different poses. In this particular experiment, the template training data contains only the transformations of subject and pose. In order to factor out other common transformations, we use the two eye-center locations for locating and aligning the face. In a future experiment, when we include additional transformations, we drop this requirement and DIKF is applied completely landmark-free (at evaluation).

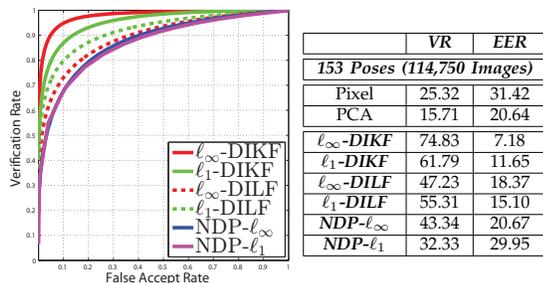


Figure 2: Pose-invariant face recognition results on the semi-synthetic large-scale mugshot database (testing on 114,750 images).



Figure 3: Samples from the testing set containing randomized synthetic transformations for one subject. (L-R): Pose, Noisy background and Scale only; Pose, Noisy background and Translation only; Pose, Noisy background and In-plane Rotation only; All transformations together⁵

benchmark against NDP- l_∞ and NDP- l_1 [26, 2] directly on *raw-pixels*. We train on 250 subjects (38,250 images) and test each method on the remaining 750 subjects (114,750 images), matching all pose-varied images of a subject to each other. Since we do not model our testing subjects during training, the protocol followed is an *Open Set (OS)* face verification protocol.

Results. Fig. 2 shows the experimental results with receiver operating characteristics (ROC) curves obtained for all methods⁴. We notice that l_∞/l_1 -DIKF performs considerably better on this task than the other methods. Therefore, in future experiments that are more competitive, we focus on DIKF. Also, we find that NDP- l_1 [26] (which was shown to perform very competitively on LFW after using some pre-processing) and NDP- l_∞ [2], are outperformed by all discriminative methods *i.e.* l_∞/l_1 -DIKF and l_∞/l_1 -DILF. We find this trend to be consistent in all our experiments. However, it is interesting to note that even vanilla dot-products followed by pooling (NDP- l_∞), perform as well as they do, suggesting that the goal of explicitly factoring out invariance has some merit. The specific choice of moment of the invariant distribution to measure appears to have an effect only in discriminative methods.

B. Open set (OS) landmark-free face recognition on large-scale semi-synthetic mugshot database using two-level sequential invariance generation. In this exploratory experiment, we simulate a controlled situation closer to real-world unconstrained face matching.

⁴Throughout this work, we use the cosine distance metric to generate the ROC curves.

Protocol. We train a two-level invariance generation framework (*i.e.* running Algorithm 1 in sequence) where the face data set at level 1 is designed to generate invariance for only pose variation (we use the same 250 subjects for pose invariance as in the previous experiment), and the one at level 2, generates invariance towards the four sets of transformations (translation, scale, in-plane rotation and all together). At level 2, we construct four distinct template sets for each of the aforementioned variations exclusively, and another one for all three transformations together. Note that to generate invariance at level 2, we need to store the pose-invariant features of the training faces from level 1 for each transformation set. We use each template set at level 2 to generate transformation invariance from the pose-invariant features obtained in level 1. The sequential training follows what is described in Section 4 and tests on the features of the randomly generated *unseen* test sets. Note that for this experiment, we are completely *landmark* and *alignment*-free at evaluation, since we match using the images in Fig. 3 as is without *any* additional information. The hope remains that the framework would be invariant towards all other transformations in the template sets while being discriminative to the transformation across subjects.

Data generation for level 2 training and overall testing. To construct the face data for training the template filter sets for each of the four transformations, we generate a 3D model of 100 training subjects, and render them at each of the 153 different poses as before. However, we now also synthetically add in the 4 sets of additional transformations (translation, scaling, in-plane rotation and all of them together). Further, we fill the background of the face wherever visible with random noise. Thus overall, we obtain 4 data sets of 15,300 images each. To generate our Open Set testing data, we repeat the randomized process to generate another four 15,300-image sets using the four sets of transformations for 100 *unseen* subjects. This randomly generated dataset (samples shown in Fig. 3) has a very challenging protocol *without any landmarks or alignment*.

Results. Fig. 4 shows the ROC curves. Level 2 l_∞/l_1 -DIKF outperforms the other methods in all cases⁶. Further, it performs surprisingly well given the difficulty in the cases of adding individual transformations to pose variation. Adding in all transformations and following the OS protocol seems to be too difficult to handle well for all methods. Roughly, in a decreasing order of difficulty the transformations are, all transformations, translation, scale followed by in-plane rotation.

C. MPIE database.

Protocol. There has not been a universal evaluation protocol established for the MPIE database. Therefore, we follow the pose-invariant face recognition protocol as adopted

⁶NDP fails to extract any discrimination among subjects at level 2 for the OS protocol.

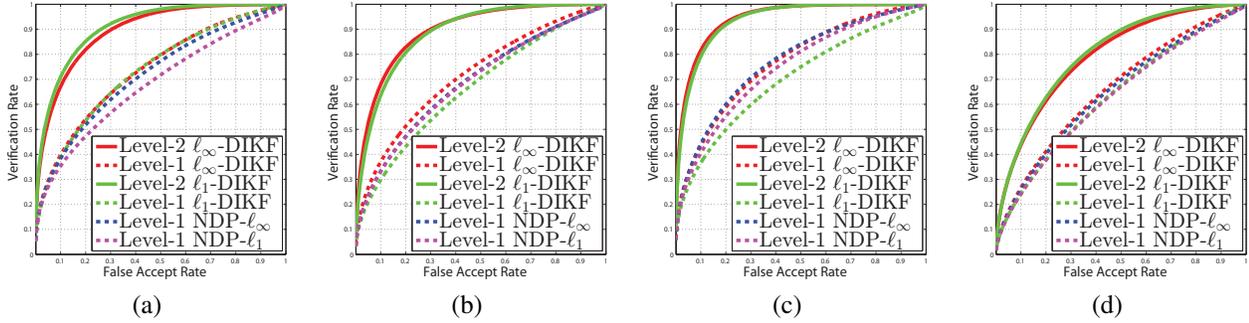


Figure 4: ROC curves for the Level-1 and Level-2 Open Set face matching protocol on large-scale mugshots database (*using no landmarks*). Different combinations of transformations are modeled for invariance (a): Pose, Noisy background and Scale only, (b): Pose, Noisy background and Translation only, (c): Pose, Noisy background and In-plane Rotation only, and (d): All transformations together.

in previous studies [31, 15, 1]. The authors use a subset of MPIE from the first session with neutral expression and frontal lighting, covering nine poses (-60° to 60° in yaw in steps of 15°) of 249 subjects⁷. We, however, generate 153 poses for each of the gallery image offline and compile our face template set. Since MPIE has 249 subjects, for each query image we form a 249 dimensional pose-invariant feature vector. We then match the features to that of the gallery image. In a parallel experiment, we also explore matching all off-angle posed images against the frontal images. This is a much harder protocol than matching each angle in yaw separately.

Results. The results are shown in Table 1. We see that l_∞ -DIKF achieves the state-of-the-art in terms of rank-1 identification rate for all poses except 60° . We attribute this to the fact that we do not generate faces beyond 40° yaw during training and so we observe graceful degradation for higher yaw angles. As an important distinction, we do not use any form of landmarking, unlike [31, 15, 1], further our method can support open-set matching, unlike [31, 15]. Fig. 5 shows the ROC curves obtained for the parallel experiment of matching all off-angle posed images against the frontal images. We find that l_∞ -DIKF significantly outperforms other methods.

D. LFW: real-world unconstrained face recognition database. Here, we apply the l_∞ -DIKF methods on a real-world unconstrained face recognition database: the LFW database [21, 19].

Protocol. We follow the standard *Unsupervised* protocol as well the *Image-Restricted, Label-Free Outside Data* protocol. Note the transformations in LFW probably do not form a group, we only focus on generating invariance towards pose in this experiment and use only the two eye-center locations for alignment. The DIKF training is carried out on *raw pixels* of a set of 3,000 randomly chosen label-free face images (which can and have multiple images from the same

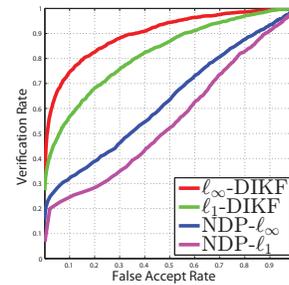


Figure 5: MPIE results on the parallel experiment: All non-frontal images matched against frontal ones.

subject). In order to abide by the *Unsupervised* protocol, we treat each face image as if they are from different subjects, making our algorithm completely agnostic about label information. Our training procedure also satisfies the second protocol mentioned above.

Results. The results for the *Unsupervised* protocol is reported in Table 2 in terms of AUC, and the mean accuracy for the *Image-Restricted, Label-Free Outside Data* protocol, along with many other competing algorithms. Fig. 6 shows the ROC curves for the LFW evaluation. The performance obtained by the proposed l_∞ -DIKF method is on par with the state-of-the-art. This is somewhat surprising considering we use only rough alignment and work directly on raw pixels unlike some of the other methods we outperform.

5.2. Dense-Landmark-Free Pose Estimation

Protocol. For the experiment on pose estimation, we generate 15 poses (-40° to 40° along yaw and -20° to 20° along pitch all in steps of 20°) for 350 *unseen* subjects (5,250 images). For this experiment, we utilize only the two eye-center locations for alignment, hence being dense-landmark-free. We construct the data set for training templates using images aligned using the eye-center locations from 250 of these subjects and their poses. Recall that the only difference between the approach for face recog-

⁷In their work, pitch variation was not explored and to have a fair comparison we restrict ourselves exclusively to variation along the yaw axis.

Table 1: Rank-1 ID rate and VR at 1% FAR (*Italics*) for ℓ_∞ -DIKF against five previous benchmarks for the MPIE database.

	L60°	L45°	L30°	L15°	R15°	R30°	R45°	R60°	All
Pixel [31]	0.4 (4.3)	0.4 (4.3)	8.8 (7.6)	15.7 (17.3)	24.9 (20.9)	9.6 (11.6)	1.2 (5.7)	0.8 (5.0)	5.6 (3.7)
PCA [31]	2.4 (3.6)	1.2 (4.3)	18.5 (9.6)	24.5 (20.1)	31.3 (23.3)	18.1 (11.6)	2.0 (7.2)	2.4 (5.0)	10.1 (7.9)
Prabhu <i>et al.</i> [31]	44.9 (26.4)	65.0 (37.1)	86.7 (59.4)	97.6 (75.5)	93.2 (71.7)	83.5 (49.0)	65.0 (45.0)	43.1 (29.7)	- (-)
Heo <i>et al.</i> [15]	- (-)	- (-)	87 (-)	96 (-)	93 (-)	90 (-)	- (-)	- (-)	- (-)
Abiantun <i>et al.</i> [1]	- (-)	51.00 (-)	85.94 (-)	97.18 (-)	97.18 (-)	87.95 (-)	53.41 (-)	- (-)	- (-)
ℓ_∞ -DIKF	27.7 (21.7)	81.9 (64.7)	92.4 (77.9)	98.8 (97.6)	99.6 (98.8)	94.4 (76.7)	83.1 (55.8)	28.1 (21.7)	75.75 (49.9)
ℓ_1 -DIKF	4.4 (0.8)	31.3 (31.3)	61.5 (60.2)	98.8 (96.8)	99.6 (99.2)	64.3 (64.3)	28.1 (36.9)	8.0 (11.2)	49.5 (36.4)
NDP- ℓ_∞	2.8 (1.6)	1.6 (2.4)	6.4 (6.0)	84.7 (75.9)	92.0 (79.5)	18.5 (21.7)	4.4 (2.0)	2.4 (1.2)	26.6 (23.2)
NDP- ℓ_1	0.4 (0.4)	0.4 (0.4)	0.8 (1.6)	64.7 (28.5)	56.6 (55.8)	3.2 (2.8)	0.4 (0.1)	0.4 (0.4)	15.9 (6.1)

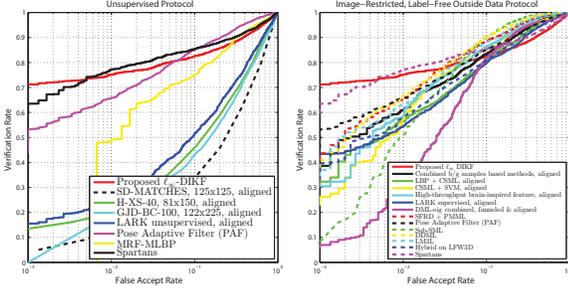
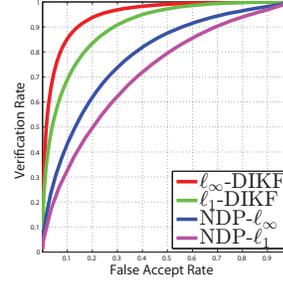


Figure 6: ROC curves for LFW evaluation under (L) Unsupervised protocol, and (R) under Image-Restricted, Label-Free Outside Data protocol.

Table 2: LFW results for the two protocols [18].

Unsupervised Protocol	
Approach	Results AUC
Proposed ℓ_∞-DIKF	0.9154
SD-MATCHES, 125 × 125 [10], aligned	0.5407
H-XS-40, 81 × 150 [10], aligned	0.7547
GJD-BC-100, 122 × 225 [10], aligned	0.7392
LARK unsupervised [33], aligned	0.7830
LHS [34], aligned	0.8107
Pose Adaptive Filter (PAF) [40]	0.9405
MRF-MLBP [4]	0.8994
MRF-Fusion-CSKDA [5] (no ROC)	0.9894
Spartans [22]	0.9428
Image-Restricted, Label-Free Outside Data Protocol	
Approach	Results $\mu \pm S_E$
Proposed ℓ_∞-DIKF	0.8867 ± 0.0046
Combined b/g samples based [39], aligned	0.8683 ± 0.0034
LBP + CSML [27], aligned	0.8557 ± 0.0052
CSML + SVM [27], aligned	0.8800 ± 0.0037
High-throughput brain-inspired [30], aligned	0.8813 ± 0.0058
LARK supervised [33], aligned	0.8510 ± 0.0059
DML-eig combined [41], funneled, aligned	0.8565 ± 0.0056
SFRD + PMML [9]	0.8935 ± 0.0050
Pose Adaptive Filter (PAF) [40]	0.8777 ± 0.0051
Convolutional DBN [20] (no ROC)	0.8777 ± 0.0062
Sub-SML [7]	0.8973 ± 0.0038
VMRS [6] (no ROC)	0.9110 ± 0.0059
DDML [16]	0.9068 ± 0.0141
LM3L [17]	0.8957 ± 0.0153
Hybrid on LFW3D [13]	0.8563 ± 0.0053
Spartans [22]	0.8969 ± 0.0036
MSBSIF-SIEDA [28]	0.9463 ± 0.0095

dition and pose estimation is that in Algorithm 1, G models pose variation for face recognition and subject variation for pose estimation (thereby being a common single framework for both tasks). We train on the 250 subjects with their poses and then test on the 1,500 images of the remaining



	VR	EER	Rank-1
ℓ_∞ -DIKF	40.70	11.71	86.47
ℓ_1 -DIKF	28.84	18.27	79.07
NDP- ℓ_∞	15.91	28.03	65.80
NDP- ℓ_1	4.67	34.22	60.40

Figure 7: Pose Estimation results on a subset of the large-scale mugshots database generated in Section 5.1 (testing on 1,500 images).

100 subjects. For each test image, we extract a subject-invariant pose-specific DIKF. We then match all features across each other and verify using the ground truth pose labels (15 poses).

Results. Fig. 7 presents the ROC curves and statistics obtained for this experiment. We find that ℓ_∞ -DIKF achieves a much better accuracy of 86% using only the two eye-center coordinates under this challenging protocol, thereby demonstrating the efficacy of the approach.

6. Conclusions

This paper presents a ‘bells and whistles’ free approach to learn or extract discriminative features that are invariant to unitary transformations. The theoretical results allow discriminative and kernelized features based on unitary kernels, to achieve group invariance through moments, thereby allowing for much more complex models to guarantee invariance. It proposes a *single* framework for face recognition and pose estimation and the practical algorithm used was always restricted to be ‘simple’, all the while working on *raw-pixels*. Yet, DIKF outperforms many previous, more complex methods and achieves (MPIE) or is on par with the state-of-the-art (LFW). The sequential generation of invariance is able to handle much more challenging protocols, unlike pooling over the traditional NDP. Although we focus only on face recognition and pose estimation, the results provide more evidence towards the hypothesis that perhaps a careful balance between invariance and selectivity is important for general vision tasks.

References

- [1] R. Abiantun, U. Prabhu, and M. Savvides. Sparse feature extraction for pose-tolerant face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014. 7, 8
- [2] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio. Magic materials: a theory of deep hierarchical architectures for learning sensory representations. *MIT, CBCL paper*, 2013. 2, 3, 5, 6
- [3] S. Arashloo and J. Kittler. Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features. *Information Forensics and Security, IEEE Transactions on*, 9(12):2100–2109, Dec 2014. 2
- [4] S. R. Arashloo and J. Kittler. Efficient processing of mrfs for unconstrained-pose face recognition. In *IEEE BTAS*, 2013. 8
- [5] S. R. Arashloo and J. Kittler. Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features. *IEEE Transactions on Information Forensics and Security*, 9(12):2100–2109, Dec 2014. 8
- [6] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz. Fast high dimensional vector multiplication face recognition. In *IEEE ICCV*, 2013. 1, 2, 5, 8
- [7] Q. Cao, Y. Ying, and P. Li. Similarity metric learning for face recognition. In *IEEE ICCV*, 2013. 2, 5, 8
- [8] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2013. 1, 2, 5
- [9] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. 8
- [10] J. R. del Solar, R. Verschae, and M. Correa. Recognition of faces in unconstrained environments: A comparative study. *EURASIP Journal on Advances in Signal Processing (Recent Advances in Biometric Systems: A Signal Processing Perspective)*, 2009. 8
- [11] C. Ding, J. Choi, D. Tao, and L. S. Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *CoRR*, abs/1401.5311, 2014. 2
- [12] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *CoRR*, abs/1509.00244, 2015. 2
- [13] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *IEEE CVPR*, 2015. 8
- [14] J. Heo. Generic elastic models for 2d pose synthesis and face recognition. *PhD Thesis*, 2009. 4
- [15] J. Heo and M. Savvides. Gender and ethnicity specific generic elastic models from a single 2d image for novel 2d pose face synthesis and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(12):2341–2350, Dec 2012. 7, 8
- [16] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *IEEE CVPR*, 2014. 8
- [17] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *Asian Conference on Computer Vision (ACCV)*, 2014. 8
- [18] G. Huang. Results on the Labeled Faces in the Wild, may 2015. 8
- [19] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. In *UMass Amherst Technical Report UM-CS-2014-003*, 2014. 7
- [20] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *IEEE CVPR*, 2012. 8
- [21] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *Technical Report of Univ. of Massachusetts, Amherst*, (07-49), oct 2007. 1, 7
- [22] F. Juefei-Xu, K. Luu, and M. Savvides. Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios. *Image Processing, IEEE Transactions on*, 24(12):4780–4795, Dec 2015. 2, 8
- [23] J. Z. Leibo, Q. Liao, and T. Poggio. Subtasks of unconstrained face recognition. *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. (VISAPP)*, 2014. 1
- [24] H. Li and G. Hua. Hierarchical-pep model for real-world face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4055–4064, June 2015. 2
- [25] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2013. 2
- [26] Q. Liao, J. Z. Leibo, and T. Poggio. Learning invariant representations and applications to face verification. *Advances in Neural Information Processing Systems (NIPS)*, 2013. 1, 2, 3, 5, 6
- [27] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision (ACCV)*, 2010. 8
- [28] A. Ouamane, M. Bengherabi, A. Hadid, and M. Cheriet. Side-information based exponential discriminant analysis for face verification in the wild. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 02, pages 1–6, 2015. 8
- [29] S. W. Park and M. Savvides. An extension of multifactor analysis for face recognition based on submanifold learning. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2645–2652, June 2010. 4
- [30] N. Pinto and D. Cox. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2011. 8
- [31] U. Prabhu, J. Heo, and M. Savvides. Unconstrained pose-invariant face recognition using 3d generic elastic models. *TPAMI*, 33(10):1952–1961, oct 2011. 2, 5, 7, 8
- [32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

- [33] H. J. Seo and P. Milanfar. Face verification using the lark representation. *IEEE Transactions on Information Forensics and Security*, 6(4):1275–1286, 2011. 8
- [34] G. Sharma, S. ul Hussain, and F. Jurie. Local higher-order statistics (lhs) for texture categorization and facial analysis. In *European Conference on Computer Vision (ECCV)*, 2012. 8
- [35] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *British Machine Vision Conference (BMVC)*, 2013. 1, 2, 5
- [36] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1988–1996. Curran Associates, Inc., 2014. 2
- [37] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1891–1898, June 2014. 2
- [38] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708, June 2014. 1, 2
- [39] L. Wolf, T. Hassner, and Y. Taigman. Similarity scores based on background samples. In *Asian Conference on Computer Vision (ACCV)*, 2009. 8
- [40] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. 1, 2, 5, 8
- [41] Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research (Special Topics on Kernel and Metric Learning)*, 2012. 8
- [42] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 787–796, June 2015. 2