Discriminative Pose-Free Descriptors for Face and Object Matching

Soubhik Sanyal, Sivaram Prasad Mudunuri and Soma Biswas
Indian Institute of Science, Bangalore, India.
{soubhiksanyal,sivaram.prasad,soma.biswas} @ee.iisc.ernet.in

Abstract

Pose invariant matching is a very important and challenging problem with various applications like recognizing faces in uncontrolled scenarios, matching objects taken from different view points, etc. In this paper, we propose a discriminative pose-free descriptor (DPFD) which can be used to match faces/objects across pose variations. Training examples at very few representative poses are used to generate virtual intermediate pose subspaces. An image or image region is then represented by a feature set obtained by projecting it on all these subspaces and a discriminative transform is applied on this feature set to make it suitable for classification tasks. Finally, this discriminative feature set is represented by a single feature vector, termed as DPFD. The DPFD of images taken from different viewpoints can be directly compared for matching. Extensive experiments on recognizing faces across pose, pose and resolution on the Multi-PIE and Surveillance Cameras Face datasets and comparisons with state-of-the-art approaches show the effectiveness of the proposed approach. Experiments on matching general objects across viewpoints show the generalizability of the proposed approach beyond faces.

1. Introduction

Matching objects across pose is a very important area of research in the field of computer vision with many applications. For example, in surveillance setting, the face of a person captured by the overhead cameras may be in any uncontrolled pose and resolution as opposed to the frontal image under high resolution that is typically captured during enrolment (Figure 1 left). For object matching, the images captured during testing can be taken from a different viewpoint compared to the images stored in the database which again requires comparing objects present in different poses (Figure 1 right).

In this paper, we propose a discriminative pose-free descriptor (DPFD) for matching objects across different poses. In the training phase, training images from few poses

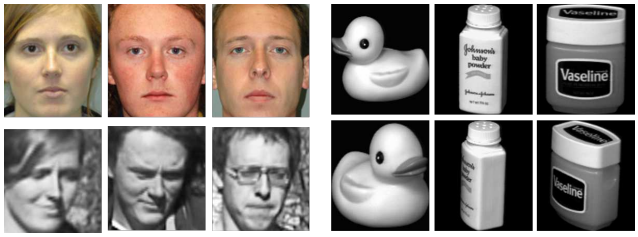

Figure 1. Applications of pose invariant matching. (Left) Face recognition in uncontrolled setting; (Right) Object recognition across viewpoint.

(two/three) are used to generate virtual subspaces for the intermediate poses. Treating the subspaces generated by the training data as points on the Grassmann manifold, intermediate subspaces are generated by sampling the shortest geodesic path between those points. An image or image region is then represented by a set of features, computed by projecting its feature vector onto all the intermediate subspaces. This will ensure that at least one or more of the features from the entire feature set will match when we try to match images with different poses. Since the final goal is recognition, a discriminative transform learned using the class labels of the training data is used to transform the feature set. Then a single compact discriminative feature vector termed discriminative pose-free descriptor (DPFD) is computed from the feature set which can be directly used for matching.

In this paper, we focus on the face recognition application where the gallery consists of frontal images captured during enrolment and the probe images can be in any uncontrolled pose. We also extend the approach when in addition to non-frontal pose, the probe images also have low-resolution as is usually the case in surveillance setting when the images are taken from a large distance from the subject. We perform extensive experiments on the CMU PIE, Multi-PIE and the Surveillance Cameras Face Database. Comparisons with state-of-the-art metric learning, cross-pose methods, domain adaptation and coupled dictionary learning approaches show the effectiveness of the proposed approach. We also show the usefulness of the proposed approach for matching general objects across pose. The nov-

elty/contribution of the proposed approach is as follows:

- A novel discriminative pose-free descriptor (DPFD) for matching objects across different poses.
- The approach does not require separate training for different probe poses/viewpoints. This is an advantage over many other approaches which work well when separate training is performed for different poses encountered during testing.
- Very few poses (as little as two/three) are required during the training phase and the method can generalize to unseen poses.

The rest of the paper is organized as follows. Section 2 describes the related literature. Details of the proposed approach is given in Section 3 and the experimental results are reported in Section 4. The paper concludes with a brief discussion section.

2. Related Work

In this section, we provide pointers to some of the related work in the area of recognizing faces and objects across pose. Handling pose variations is a commonly addressed issue in matching facial images [22][37][20][24]. Kan *et al.* [19] learn a model to transform the complex non-linearity of the non-frontal facial images to frontal ones with a deep network. Xiong and Torre [36] propose a face alignment algorithm that uses supervised descent method which learns a sequence of descent directions required to minimize the mean of the non-linear least squares function. An automatic 3D pose normalization approach that can synthesize frontal view of every gallery and pose images by fitting a 3D face model to a 2D input facial image is described in [3]. Ashraf *et al.* [2] propose a framework to align facial images of different views at patch level and matching is performed by using the discriminative power of the corresponding gallery and probe patches. Castillo and Jacobs [8] propose a method to compute stereo matching cost between two facial images by using epipolar geometry.

Recently, matching of low resolution facial images has gained considerable attention [16][25][23][31][27]. Biswas *et al.* [7] suggest a method of using multidimensional scaling to transform the features from low resolution (LR) non-frontal probe images and the high resolution (HR) frontal gallery images to a common space. A piecewise linear regression model is developed to learn the relationship between the HR image space and the LR image space for face super resolution in [38] . Jiang *et al.* [18] propose to super-resolve the HR version of a LR probe image by manifold learning and discriminant analysis and then perform recognition. A framework of co-transfer learning as a mixture of transfer learning and co-training paradigms for matching faces can be found in [6].

Metric learning approaches have shown a lot of promise for matching faces in unconstrained environments. Kostinger *et al.* [21] propose a method that learns a distance metric from the co-variance matrices of similar and dissimilar pairs. Moutafis and Kakadiaris [27] propose an algorithm that can match HR and LR facial images by learning individual basis for optimal representation and coupled distance metrics to enhance the classification. Domain adaptation techniques have also been successfully used for matching face images across pose, pose and blur, etc. [11]. In [29], dictionary learning is used to interpolate subspaces to link the source and target domains.

There has also been a lot of research in the area of general object recognition across different viewpoints [32]. Ozay *et al.* [30] propose a joint object pose estimation and categorization approach by constructing a hierarchical object representation and extracting information from the object parts and compositions from different layers of the hierarchy. A model that separates a view-invariant category representation from category-invariant pose representation is proposed in [5]. Aytar and Zisserman [4] propose a convex optimization based model transfer learning algorithm to categorize the objects.

Our work is inspired by [15] in which images are matched across varying scales. Features at different scales can be computed from the same image itself, unlike features at different poses which is the focus of our work. Generating intermediate subspaces by sampling the Grassmann manifold has also been exploited by [11], and then the projections on these subspaces are used to train discriminative classifiers for each object. Instead, using the intermediate subspaces, we form a discriminative feature vector which can directly be used for matching. Our approach is more suitable for applications like face recognition, where there may not be any overlap between the training and testing subjects.

3. Proposed Approach

Here we describe in detail how to construct the discriminative pose-free descriptor (DPFD). In the training phase, given training examples from a few pose regions, virtual intermediate subspaces are created. The feature vector from the input image (or image region) is projected onto all the subspaces to form a feature set. A discriminative transform is then learnt from the training class labels. In the testing phase, after computing the feature set for a given image by projecting onto all the subspaces, they are transformed using the discriminative transform learnt in the training phase. A single feature vector is finally constructed from the feature set which is the discriminative pose-free descriptor (DPFD) for the image. The flowchart of the proposed approach is shown in Figure 2. We describe each of the steps in detail in the following subsections.

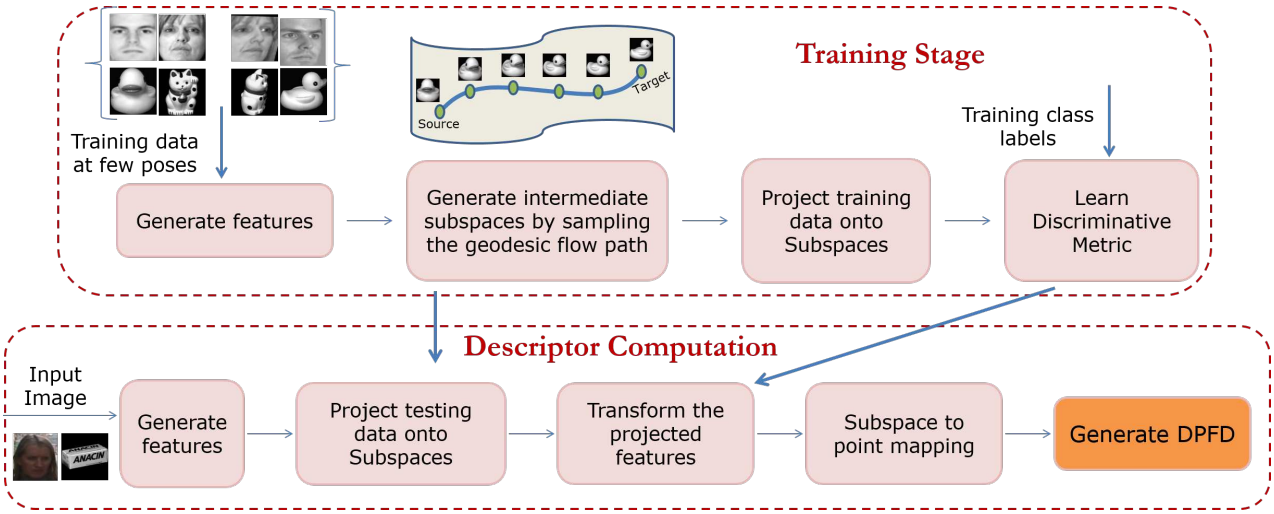

Figure 2. Flow chart of the proposed algorithm showing the training stage and construction of the DPFD.

3.1. Feature Representation using Intermediate Subspaces

Suppose we have training images from some parts of the pose space, say, P_1, P_2 to P_K, (K is as small as two/three) (Figure 4). The goal is to compute a descriptor for an image with any unknown pose so that it can be used for matching across poses. Let the feature descriptor of an image (or image region) be denoted as f. Since the actual image to be matched can be in any pose, instead of representing the image by f, we propose to represent it using a collection of features $\{f_1, f_2, \ldots\}$, which are the feature vectors computed if we have the image at different poses. If we now compare the feature sets from two images of the same object which differ by pose, it is more likely that one or more of the features from the sets will match, as compared to if we had represented the images with only f.

In order to generate the features at different poses, we compute virtual poses by learning the path between P_k and P_{k+1}. To do so, we exploit the idea of sampling on the Grassmann manifold [11][9]. Let the features corresponding to the images in P_k be denoted as $f_{k,i} \in \mathbb{R}^D$, where $i = 1, 2, \ldots N$ denote the training images in pose P_k and similarly for pose P_{k+1} Thus, we have a data matrix of dimension $D \times N$ for pose P_k as well as P_{k+1}. Let $\bar{P}_k$ and $\bar{P}_{k+1} \in \mathbb{R}^{(D \times d)}$ be the corresponding generative subspaces obtained by applying Principal Component Analysis (PCA) on the data matrix P_k and P_{k+1} respectively. The space of d-dimensional subspaces in $\mathbb{R}^D$ can be identified with the Grassmann manifold $\mathbb{G}_{d,D}$ and thus, $\bar{P}_k$ and $\bar{P}_{k+1}$ are points on $\mathbb{G}_{d,D}$. Our goal is to obtain the intermediate subspaces between $\bar{P}_k$ and $\bar{P}_{k+1}$ and generate the virtual features corresponding to those subspaces.

Let $R_k \in \mathbb{R}^{D \times (D-d)}$ represent the orthogonal complement of $\bar{P}_k$, which implies $R_k^T \bar{P}_k = 0$. The geodesic flow between $\bar{P}_k$ and $\bar{P}_{k+1}$ is given by $\Psi(t) : t \in [0, 1]$, such that $\Psi(t) \in \mathbb{G}_{d,D}$ and $\Psi(0) = \bar{P}_k$ and $\Psi(1) = \bar{P}_{k+1}$, i.e., the geodesic flow starts from $\bar{P}_k$ and reaches $\bar{P}_{k+1}$ in unit time. The expression for the flow at any time t is given by

$$\Psi(t) = \bar{P}_k U_1 \Gamma(t) - R_k U_2 \Sigma(t) \tag{1}$$

where $U_1 \in \mathbb{R}^{d \times d}$ and $U_2 \in \mathbb{R}^{(D-d) \times d}$ are orthonormal matrices given by $\bar{P}_k^T \bar{P}_{k+1} = U_1 \Gamma V^T$ and $R_k^T \bar{P}_{k+1} = -U_2 \Sigma V^T$. Γ and Σ are $d \times d$ diagonal matrices whose diagonal elements are $cos\theta_i$ and $sin\theta_i$ for $i = 1, 2, , \ldots d$. θ_i are known as the principal angles between $\bar{P}_k$ and $\bar{P}_{k+1}$. $\Gamma(t)$ and $\Sigma(t)$ are diagonal matrices whose elements are $cos(t\theta_i)$ and $sin(t\theta_i)$ respectively. The different intermediate subspaces are obtained for different values of t. The idea behind our approach is that, if we project an image in any unknown pose on any of the interpolated subspaces, the reconstructed image will have a pose similar to that of the interpolated pose. We have projected original images (Figure 3a) of frontal (top row) and 30° pose (bottom row) onto an interpolated pose (15°). We observe that the reconstructed images (b) are close to the actual 15° pose (c) in both cases thus justifying the subspace interpolation.

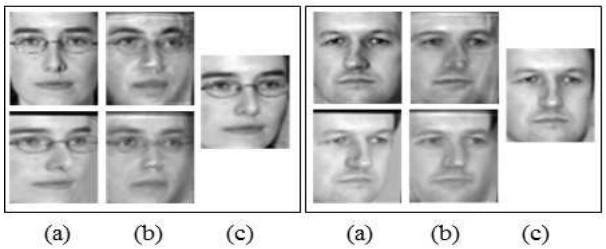

Figure 3. Illustration of Pose reconstruction on Geodesic flow curve for two subjects. a: unknown pose; b: synthesized from interpolated subspace; c: actual image at interpolated pose.

Figure 4. Top: Training images from 3 different parts of the pose space (left pose, frontal, right pose) denoted by P_1, P_2 and P_3 respectively. Bottom: Virtual subspaces generated from training data, the two rows indicating the second and fourth eigenvectors of the subspaces.

After getting the intermediate subspaces $\Psi(t)$, the feature vector f is projected onto all the subspaces obtaining the enhanced feature set for each image $\{f_1, f_2, \dots f_N\}$. Here N is the total number of subspaces, out of which K subspaces are computed from the actual training data and $(N-K)$ are the intermediate virtual subspaces.

Figure 4 (bottom) shows virtual subspaces generated from training data from three parts of the pose space (Figure 4 (top)), the two rows indicating the second and fourth eigenvectors of the subspaces.

To illustrate the effectiveness of the proposed feature set representation over the standard feature vector representation for matching across poses, we perform an experiment on the Multi-PIE dataset [14] using images of 100 subjects, under frontal illumination condition and five different poses including the frontal pose (Figure 5). The corner of the right eye is represented using the SIFT descriptor and also using the proposed feature set representation. Each point in the Figure 5 shows the difference between the descriptor at that pose from the frontal image, averaged over all the 100 subjects. For computing the difference between two feature sets for our descriptor, we computed the distances between all the features and took the minimum. We see that the difference increases as the difference in pose increases, but the difference is much less for the proposed descriptor as opposed to the baseline SIFT descriptor, indicating that the proposed descriptor is more robust to change in pose.

But there are two issues that need to be addressed before such a feature set can be used for matching across pose variations

1. The feature sets have been computed from generative subspaces, so they may not be discriminative enough for recognition/classification task.

2. If two feature sets are matched using some measure like minimum distance, then $N \times N$ comparisons are required to compute the distance between two feature sets, which is computationally expensive.

We address both the issues in the following subsections.

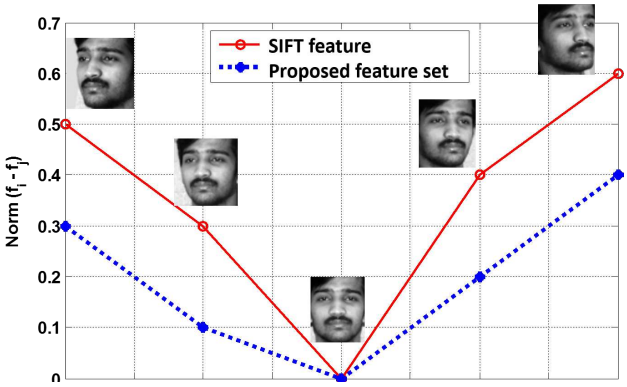

Figure 5. Proposed feature set representation generated using projections on all intermediate virtual subspaces vs. the SIFT descriptor. Each point in the curves indicate the difference of the feature vector for that pose from that computed from the frontal image.

3.2. Discriminative Features

In this section, we describe how to construct a discriminative feature set for a given input image from the feature set that is computed from generative subspaces to apply for the task of recognition, which is our final objective. The class labels of the training data are utilized to learn a transformation such that transformed feature sets of the same class come closer to one another and those of different classes are moved further away. In our work, we use the framework of Mahalanobis distance metric learning to make features discriminative. In general, the squared distance between two features x_i, x_j can be defined as

$$d^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j) \qquad (2)$$

where $M \succeq 0$ is the positive semi-definite matrix that we want to learn. Since the difference in pose between the images to be matched may be significant (considering the two extremes of the pose space as in Figure 4), one metric may not be sufficient. So, the whole pose space is divided into say T regions and we propose to learn a metric for each of these regions. For example, if there are 12 subspaces and 4 regions, each region will constitute of 3 subspaces. For each

region, the features for the constituent subspaces are concatenated and used as input feature vector for Mahalanobis metric learning. Here, we have utilized a formulation similar to the Large Scale Metric Learning (LSML) [21] for learning the metrics for each of these T regions. We provide some details of the LSML algorithm for completeness.

The approach considers two independent generation processes for match and non-match pairs. The features that are from same subject which has variations in pose, illumination and resolution are the match pairs. The features that are from different subjects are the non-match pairs. For any given pair of features x_i and x_j of a region, the decision on whether they belong to same class or not can be obtained from likelihood ratio test as formulated below

$$\delta(x_i, x_j) = log\left(\frac{p(x_i, x_j)|H_0)}{p(x_i, x_j)|H_1}\right) \tag{3}$$

where, H_0 and H_1 are the hypotheses that a pair is non-match and match respectively. If a pair of features belong to the same class, the value of $\delta(x_i, x_j)$ is small, otherwise it is large.

The problem in (3) can be reformulated by assuming a Gaussian structure for the difference space of features as

$$\delta(x_{ij}) = log\left(\frac{\frac{1}{\sqrt{2\pi|\Sigma_{n_{ij}=0}|}}exp\left(-\frac{1}{2}x_{ij}^T\Sigma_{n_{ij}=0}^{-1}x_{ij}\right)}{\frac{1}{\sqrt{2\pi|\Sigma_{n_{ij}=1}|}}exp\left(-\frac{1}{2}x_{ij}^T\Sigma_{n_{ij}=1}^{-1}x_{ij}\right)}\right) \tag{4}$$

where, $x_{ij} = x_i - x_j$ is a vector in the difference space; $n_{ij} = 1$ for a match pair and its value is 0 for a non-match pair. $\Sigma_{n_{ij}=1}$ and $\Sigma_{n_{ij}=0}$ are the corresponding covariance matrices. The above equation can be simplified as

$$\delta(x_{ij}) = x_{ij}^T\left(\Sigma_{n_{ij}=1}^{-1} - \Sigma_{n_{ij}=0}^{-1}\right)x_{ij} \tag{5}$$

Analyzing (2) and (5), the Mahalanobis Metric can be given by $M = \left(\Sigma_{n_{ij}=1}^{-1} - \Sigma_{n_{ij}=0}^{-1}\right)$. Please refer to [21] for further details.

Figure 6 shows the match and non-match scores distributions before (top) and after (bottom) the discriminative transform for 2000 test image pairs of the Multi-PIE dataset. We see that the discriminative transform leads to better separation of the two distributions which results in better recognition performance.

3.3. DPFD Computation

The discriminative feature sets from two images can potentially be compared using a suitable set comparison metric to compute the distance between them. But as discussed before, this is computationally inefficient. Here we describe how to generate DPFD vector from the discriminative feature set which can be used to efficiently match two feature sets. Since we have generated virtual views between the different poses, the feature vectors at the different poses change gradually. So the set of descriptors corresponding to each region in pose space can be approximated to lie on a linear subspace. Suppose the basis vectors for region t spanning the space of the features is given by $g_{t,1}, g_{t,2}, \ldots, g_{t,N_s}$. The $D \times N_s$ matrix $G_t = [g_{t,1}, g_{t,2}, \ldots, g_{t,N_s}]$ represents the subspace for region t, where N_s is the number of basis vectors of the subspace, and the dimension of each feature vector is D. Now we compute the subspace to vector representation for each region.

The vector representation can be achieved by rearranging the elements of the $D \times D$ matrix $L = G_t G_t^T$ using the following operator (considering only the elements of the upper triangular matrix with the diagonal elements scaled by $1/\sqrt{2}$) [15]

$$\text{DPFD}_\text{t} = \left(\frac{l_{11}}{\sqrt{2}}, l_{12}, ..., l_{1D}, \frac{l_{22}}{\sqrt{2}}, l_{23}, ..., \frac{l_{DD}}{\sqrt{2}}\right)^T \tag{6}$$

Here $D = 128$, since we have used SIFT descriptor as the input feature. Finally, the vector representation for all the T regions are concatenated into a single vector termed as the DPFD given by

$$\text{DPFD} = [\text{DPFD}_1; \text{DPFD}_2; \ldots; \text{DPFD}_\text{T}] \tag{7}$$

4. Experimental Evaluation

In this section, we report the results of extensive experiments performed to evaluate the effectiveness of the proposed approach. Specifically, we perform experiments on face recognition across pose, face recognition across pose and resolution, and object recognition across pose.

4.1. Face Recognition Across Pose

We have represented the facial images by local feature descriptors (SIFT in this paper) computed at 15 fudicial locations of face images. A freely available C++ software library based on active shape models known as STASM [26] is used to detect the fiducial locations automatically. The locations are manually verified and incorrect ones are corrected. Here, we present experiments on recognizing faces across pose variations on the CMU-PIE dataset [34]. We followed the same protocol as in [29] and used all the 68 subjects under 5 different poses and frontal illumination for this experiment. For this experiment, we have used 100 subjects from the Multi-PIE data [14] whose images have been captured under very similar conditions as the PIE data for training. We have constructed subspaces for each fiducial point separately and have used frontal and extreme poses (c_{11} and c_{37}) for representing the entire pose space during training. We have computed 12 subspaces in between pose

Table 1. Rank-1 recognition accuracies (%) for face recognition across pose variations on the PIE dataset [34].

Method	c_{11}	c_{29}	c_{05}	c_{37}	Average
K-SVD [1]	48.5	76.5	80.9	57.4	65.8
Eigen Light-field [13]	78.0	91.0	93.0	89.0	87.8
SGF [11]	58.8	89.7	89.7	72.1	77.6
GFK [10]	63.2	92.7	92.7	76.5	81.3
Subspace Interp. via DL [29]	76.5	98.5	98.5	88.2	90.4
Proposed Approach (DPFD)	**98.5**	**100**	**100**	**98.5**	**99.3**

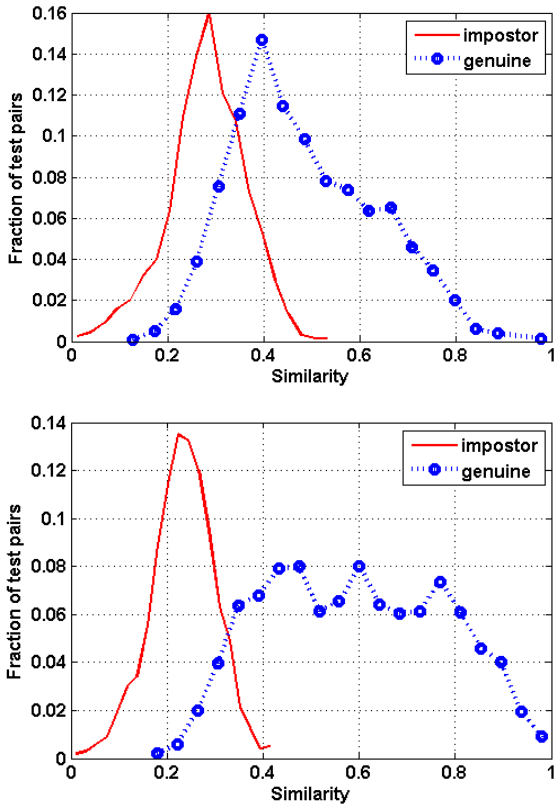

Figure 6. Match (also termed genuine) and non-match (also termed imposter) scores distributions before (top) and after (bottom) the discriminative transform. The distributions are better separated after the transformation resulting in better recognition performance.

c_{11} to frontal and frontal to pose c_{37} and the entire pose space is subdivided into 4 regions for computing the discriminative feature sets.

The frontal images are used as the gallery and the non-frontal images under the different poses are used as the probe images. There is no overlap between the subjects used in the training and testing set. Thus, there is no need for retraining even if the test subjects change. The subspaces and the transformation matrices can be learnt once offline, which can then be used for any test subject. During testing, both the gallery and the probe images are projected onto all the subspaces, their discriminative features sets are computed using the learnt metric and then the DPFDs are constructed. The DPFDs are constructed in the same way for both gallery and probe images of every pose, i.e. we do not need to learn a classifier separately for each pose.

The results of the proposed approach for this experiment is reported in Table 1. Comparison with several other approaches are shown, namely (1) K-SVD [1]: here the dictionary is learnt from the frontal images and the same dictionary is used to get the sparse coefficients for the non-frontal images; (2) SGF [11] and GFK [10]: here subspace interpolation is done on the Grassmann manifold; (3) Eigen-field approach [13] which is designed specifically to recognize faces across pose; (4) Subspace Interpolation via Dictionary Learning [29] where dictionary learning is used to interpolate subspaces to link the frontal and non-frontal domains. The recognition accuracies of all the other approaches are taken directly from [29]. We see that the proposed approach performs significantly better than all the other approaches for the task of recognizing faces across pose variations, even with no separate training for each of the different probe poses.

4.2. Face Recognition Across Pose and Resolution

The proposed approach can also be extended to recognize objects across multiple variations simultaneously. Here we show results on face recognition where the gallery is of frontal and high resolution (HR) images, while the probe images are non-frontal and of low-resolution (LR), as usually obtained from surveillance cameras.

Results on MultiPIE dataset: We report results on the Multi-PIE dataset [14] which contains images of 337 subjects from four different recording sessions captured under different poses, illumination conditions and expressions. For our experiments, we use HR images under frontal pose and frontal illumination condition as gallery. LR images taken under pose 04_1, 05_0, 13_0 and 14_0 (named as indicated in the dataset) under all the 20 different illumination conditions and neutral expression are used as the probe images. Figure 7 (a) shows some sample HR gallery and (b,c,d,e) shows the probe images under the four different probe poses. HR images of size 60×50 and LR images of

Table 2. Rank-1 recognition performance (%) for four different probe poses, averaged over the different gallery illuminations on the Multi-PIE dataset [14].

Method	Pose 13_0	Pose 14_0	Pose 05_0	Pose 04_1
MDS Learning [7]	32.8	44.8	47.0	48.5
LSML [21]	46.9	53.9	55.2	54.3
GMA [33]	65.0	70.1	70.3	64.2
SCDL [35]	66.3	73.0	72.7	64.1
CFDL [17]	65.9	72.0	72.8	64.7
Proposed (PFD)	**65.9**	**71.2**	**64.2**	**56.4**
Proposed (DPFD)	**74.5**	**78.0**	**74.0**	**70.1**

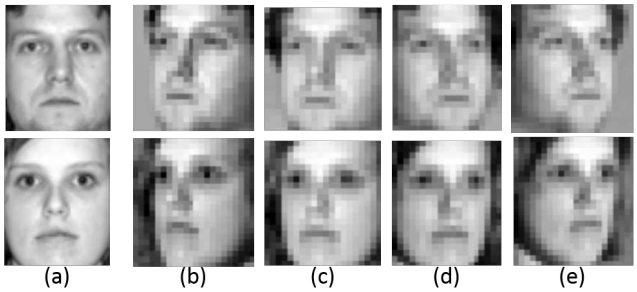

Figure 7. Example images from the Multi-PIE data [14]. (a) Frontal high-resolution images used as gallery; (b,c,d,e) low-resolution images under non-frontal pose (pose 13_0, 14_0, 05_0 and 04_1 as given in the dataset) used as probe images.

size 20×17 (i.e. scale factor of 3) are used for all the experiments. The LR images are obtained by down-sampling the original resolution images to lower resolutions using standard bi-cubic interpolation technique. We have used 100 randomly chosen subjects with frontal, 13_0 (left extreme) and 04_1 (right extreme) poses for generating the subspaces and metric learning, and the remaining subjects for testing. During subspace generation, the frontal images were of high resolution and the images for the extreme poses were of low resolution. There is no overlap between the train and test subjects. The parameters for the proposed approach are the same as used in the PIE experiment.

The results for the proposed approach are shown in Table 2. We have also compared with several state-of-the-art approaches; namely (1) MDS Transformation Learning [7] which learns a transformation between the HR frontal gallery and LR non-frontal probe; (2) Metric Learning approaches: Large Scale Metric Learning (LSML) [21] learns a metric from equivalence constraints based on the statistical inference perspective; (3) Semi-coupled and Coupled Dictionary Learning [35][17] learns the dictionaries jointly for matching objects from different domains, and (4) Generalized Multiview Analysis (GMA) [33] computes a single linear subspace by solving a joint, quadratic program over different feature spaces. PFD refers to our algorithm without the discriminative transform, i.e. the feature set computed by projecting onto all the virtual intermediate subspaces is directly converted to a single feature vector. For all the algorithms, the same input features have been given, and we have learnt one transformation for all the probe poses. The codes for the other approaches have been taken from the respective author's websites. We observe that for all the poses, the proposed approach performs significantly better than the other approaches. The observation that DPFD performs much better than PFD indicates the usefulness of the discriminative learning step in the proposed approach.

Results on Surveillance Cameras Face Database: Now we evaluate the proposed approach on real surveillance quality data obtained from the Surveillance Cameras Face Database [12]. The dataset contains images of 130 subjects captured in uncontrolled environment using five different video surveillance cameras, and the gallery images were taken using high-quality camera. We use the same experimental setup as used in [7], in which all the images from the five surveillance cameras i.e. a total of 650 images are used for the experiment. Figure 8 shows some gallery (top row) and probe images (bottom row).

Figure 8. Example facial images of Surveillance Cameras Face Database [12]. Top row: frontal gallery images, second row: corresponding probe images captured by surveillance cameras.

As in [7], we randomly pick 50 subjects for training and use the remaining 80 subjects for testing (thus there are a total of 400 probe images) with no overlap between the train and test subjects. The experiment is repeated 10 times with different random sampling of the subjects. The Rank-1 accuracy of the proposed approach and comparisons with several other approaches for this experiment are shown in Table 3. For our approach, we have used HR frontal images

Table 3. Rank-1 accuracy (%) of the proposed approach and comparison with state-of-the-art approaches on the Surveillance Cameras Face Database [12]. The two columns indicate two different training setups using data from only one camera and five cameras for training. The proposed approach trained using data from just one camera performs better than all the compared approaches even when they are trained using data from all five cameras.

Method	Rank-1 1 Cam	Rank-1 5 Cam
MDS Learning [7]	30.0	61.1
LSML [21]	64.7	67.2
GMA [33]	38.2	50.5
SCDL [35]	48.2	58.5
CFDL [17]	45.7	62.2
Proposed (PFD)	**46.0**	–
Proposed (DPFD)	**69.0**	–

and LR non-frontal images from one camera to generate the subspaces and transformation learning. For all the other approaches, we used two setups for training: (a) HR frontal images and non-frontal images from one camera (same as for the proposed approach) (Table 3 second column); (b) HR frontal images and non-frontal images from all the five cameras (Table 3 third column). We observe that when only one camera is used for training, the proposed approach performs significantly better than the other approaches. When the training of the other approaches uses images from all the cameras, their performance improves, but it is still worse than that of the proposed approach. This shows that our approach can generalize better across unseen poses.

4.3. Object Recognition Across Pose

Though all the experiments so far has been performed on facial images, the proposed approach is general can be used to recognize general objects across variations in viewpoint. Here we perform experiments on Columbia Object Image Library 20 (COIL 20) database [28]. The dataset contains 20 objects and each object is captured by rotating it about its vertical axis at a regular interval of $5°$. We selected 50 images of each object that has pose variations from left extreme to right extreme including the frontal pose for our experiments. A few sample images are shown in Figure 9. Five images per object around the frontal pose and extreme pose are used for training and the remaining images are used for testing. All the images are grayscale and are resized to 32×32 and the image intensity values are used as the input features. A total of 12 subspaces with four regions in pose space are used in this experiment.

During testing, images of frontal pose are taken as gallery data and the remaining images that differ in pose are used as probe images. The images and the poses used for testing is different from the ones used for generating the subspaces and metric learning. Note that the protocol used in our experiment is different from the ones normally used for testing object categorization algorithms. Our protocol is designed to see whether the proposed descriptor can be used for describing general objects across pose. So the performance reported cannot be directly compared with other published papers which have reported results on this dataset. Table 4 shows the Rank-1 recognition accuracy of the proposed approach and comparisons with other approaches. We see that the proposed approach performs better than all the other approaches.

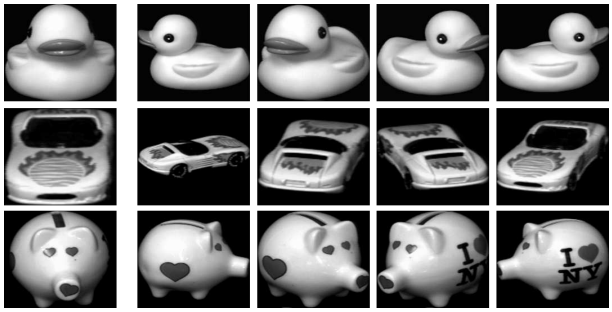

Figure 9. Sample images from the COIL 20 dataset [28]. The first column shows the gallery images and the second to fifth columns shows some probe images for the same objects.

Table 4. Rank-1 accuracy (%) of the proposed approach and comparison with other approaches on COIL 20 Database [28].

Method	Rank-1 Accuracy
MDS Learning [7]	75.6
LSML [21]	80.3
GMA [33]	66.1
SCDL [35]	79.2
CFDL [17]	78.7
Proposed (PFD)	**67.4**
Proposed (DPFD)	**82.2**

5. Discussion

In this work, we proposed a novel discriminative pose-free descriptor (DPFD) for matching objects across pose. The proposed approach requires images from a few regions of the pose space for training and does not require separate training for each probe pose. Experimental evaluations for various tasks like face recognition across pose, face recognition across resolution and pose and object recognition across different viewpoints, are conducted to evaluate the usefulness and generalizability of the approach. Superior performance of the proposed approach as compared to several state-of-the-art approaches shows the effectiveness of the approach.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse

representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] A. B. Ashraf, S. Lucey, and T. Chen. Learning patch correspondences for improved viewpoint invariant face recognition. *CVPR*, pages 1–8, 2008.

[3] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. *ICCV*, pages 937–944, 2011.

[4] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. *ICCV*, pages 2252–2259, 2011.

[5] A. Bakry and A. Elgammal. Untangling object-view manifold for multiview recognition and pose estimation. *ECCV*, 2014.

[6] H. S. Bhatt, R. Singh, M. Vatsa, and N. K. Ratha. Improving cross-resolution face matching using ensemble based co-transfer learning. *IEEE Transactions on Image Processing*, 23(12):5654–5669, 2014.

[7] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer. Pose-robust recognition of low-resolution face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3037–3049, 2013.

[8] C. D. Castillo and D. W. Jacobs. Using stereo matching with general epipolar geometry for 2d face recognition across pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2298–2304, 2009.

[9] K. Gallivan, A. Srivastava, X. Liu, and P. V. Dooren. Efficient algorithms for inferences on grassmann manifolds. *IEEE Workshop on Statistical Signal Processing*, pages 315–318, 2003.

[10] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. *CVPR*, pages 2066–2073, 2012.

[11] R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2288–2302, 2014.

[12] M. Grgic, K. Delac, and S. Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

[13] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):449–465, 2004.

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Guide to the cmu multi-pie database. *Technical report - Carnegie Mellon University*, 2007.

[15] T. Hassner, V. Mayzels, and L. Z. Manor. On sifts and their scales. *CVPR*, pages 808–821, 2012.

[16] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. *CVPR*, pages 1–8, 2008.

[17] D. A. Huang and Y. C. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. *ICCV*, pages 2496–2503, 2013.

[18] J. Jiang, R. Hu, Z. Han, K. Huang, and T. Lu. Graph discriminant analysis on multi-manifold (gdamm): a novel super-resolution method for face recognition. *ICIP*, pages 1465–1468, 2012.

[19] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. *CVPR*, pages 1883–1890, 2014.

[20] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *ECCV*, pages 808–821, 2012.

[21] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *CVPR*, pages 2228–2295, 2012.

[22] A. Li, S. Shan, and W. Gao. Coupled biasvariance tradeoff for cross-pose face recognition. *IEEE Transactions on Image Processing*, 21(1):305–315, 2012.

[23] B. Li, H. Chang, S. Shan, and X. Chen. Low-resolution face recognition via coupled locality preserving mappings. *CVPR*, pages 20–23, 2010.

[24] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. *CVPR*, pages 3499–3506, 2013.

[25] C. Liu, H. Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *CVPR*, pages 115–134, 2007.

[26] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *ECCV, http://www.milbo.users.sonic.net/stasm*, 2008.

[27] P. Moutafis and I. A. Kakadiaris. Semi-coupled basis and distance metric learning for cross-domain matching: Application to low-resolution face recognition. *IJCB*, pages 1–8, 2014.

[28] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). *Technical Report*, 1996.

[29] J. Ni, Q. Qiu, and R. Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. *CVPR*, 2013.

[30] M. Ozay, K. Walas, and A. Leonardis. A hierarchical approach for joint multi-view object pose estimation and categorization. *ICRA*, 2014.

[31] C. Ren, D. Dai, K. Huang, and Z. Lai. Transfer learning of structured representation for face recognition. *IEEE Transactions on Image Processing*, 23(12):5440–5454, 2014.

[32] J. Schels, J. Liebelt, and R. Lienhart. Learning an object class representation on a continuous viewsphere. *CVPR*, 2012.

[33] A. Sharma, A. Kumar, H. Daume, and D. Jacobs. Generalized multiview analysis: A discriminative latent space. *ICCV*, pages 2160–2167, 2012.

[34] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[35] S. Wang, D. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. *CVPR*, pages 2216–2223, 2012.

[36] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. *CVPR*, pages 532–539, 2013.

[37] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu. Random faces guided sparse many-to-one encoder for pose-invariant face recognition. *ICCV*, pages 2416–2423, 2013.

[38] W. W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2012.