Pose-Robust Recognition of Low-Resolution Face Images

Soma Biswas, *Member*, *IEEE*, Gaurav Aggarwal, *Member*, *IEEE*, Patrick J. Flynn, *Fellow*, *IEEE*, and Kevin W. Bowyer, *Fellow*, *IEEE*

Abstract—Face images captured by surveillance cameras usually have poor resolution in addition to uncontrolled poses and illumination conditions, all of which adversely affect the performance of face matching algorithms. In this paper, we develop a completely automatic, novel approach for matching surveillance quality facial images to high-resolution images in frontal pose, which are often available during enrollment. The proposed approach uses multidimensional scaling to simultaneously transform the features from the poor quality probe images and the high-quality gallery images in such a manner that the distances between them approximate the distances had the probe images been captured in the same conditions as the gallery images. Tensor analysis is used for facial landmark localization in the low-resolution uncontrolled probe images for computing the features. Thorough evaluation on the Multi-PIE dataset [1] and comparisons with state-of-the-art super-resolution and classifier-based approaches are performed to illustrate the usefulness of the proposed approach. Experiments on surveillance imagery further signify the applicability of the framework. We also show the usefulness of the proposed approach for the application of tracking and recognition in surveillance videos.

Index Terms—Face recognition, low-resolution matching, multidimensional scaling, iterative majorization

1 Introduction

Face images captured by surveillance cameras usually have poor resolution in addition to uncontrolled poses and illumination conditions that adversely affect the performance of face recognition (FR) algorithms. Traditionally, research in the area of FR has been concentrated on recognizing faces across changes in illumination and pose [2], [3], [4], [5], but there is growing interest in handling poor resolution facial images [6], [7], [8], [9], [10], [11], [12], [13]. The difference in resolution in addition to pose and illumination variations adds to the complexity of the task and limited attention has been given to addressing all these variations jointly.

Most of the existing work that addresses the problem of matching faces across changes in pose and illumination cannot be applied when the gallery and probe images are of different resolutions. The commonly used approach for matching a low-resolution (LR) probe image with a high-resolution (HR) gallery is to use super-resolution (SR) to construct a higher resolution image from the probe image and then perform matching. But the primary goal of SR approaches is to obtain a good visual reconstruction and they are usually not designed from a recognition perspective. To this end, there have been a few recent efforts that address recognition and SR simultaneously [8]. But most of them assume that the probe and gallery images are in the same pose, making them not directly applicable for more general scenarios. We build on the success of these initial efforts and propose an approach to match LR probe images taken under uncontrolled pose and illumination conditions with HR gallery images in frontal pose.

FR algorithms perform best when the gallery and probe images have the same resolution and are taken under similar (controlled) imaging conditions. Based on this intuition, we propose a multidimensional scaling (MDS) [14] based approach to transform the features from LR nonfrontal probe images and the HR frontal gallery images to a common space in such a manner that the distances between them approximate the distances had the probe images been of the same resolution and pose as the gallery images. The desired transformation is learned from the training images using the iterative majorization algorithm. The matching process involves transforming the extracted features using the learned transformation followed by euclidean distance computation. We use SIFT-based descriptors at fiducial locations on the face image as the input feature. Locating the facial landmarks in LR, nonfrontal images under varying illumination conditions is by itself a challenging task. In this work, we propose a tensor analysis-based approach to estimate approximate pose and rough locations of the facial landmarks.

Extensive experimental evaluation on the Multi-PIE dataset [1] is performed to evaluate the proposed approach. Comparisons with state-of-the-art SR [15] and classifier-based [16] approaches are performed to illustrate the usefulness of the proposed approach. Experiments on real surveillance images from the Surveillance Cameras Face Database [17] signify the applicability of the framework for matching LR images in uncontrolled pose and illumination conditions. We further show the usefulness of the proposed

• *The authors are with the Computer Vision Research Laboratory, Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556. E-mail: {sbiswas, gaggarwa, flynn, kwb}@nd.edu.*

Manuscript received 15 Nov. 2011; revised 24 June 2012; accepted 19 Mar. 2013; published online 3 Apr. 2013.
Recommended for acceptance by S. Sarkar.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-11-0825.
Digital Object Identifier no. 10.1109/TPAMI.2013.68.

approach for the application of tracking and recognition in surveillance videos.

The rest of the paper is organized as follows: An overview of the related approaches is discussed in Section 2. A brief description of the feature representation is provided in Section 3. The details of the proposed approach are provided in Section 4. Section 5 discusses the approach for automatically detecting the fiducial landmarks. The results of experimental evaluation are presented in Section 6. The details of applying the proposed approach for the task of automatic tracking and recognition in surveillance videos are provided in Section 7. The paper concludes with a brief summary and discussion. A preliminary version of this work appeared in [18] and [19].

2 Previous Work

Several approaches have been proposed in the literature for handling one or more of the factors, like pose, illumination, and resolution, which affect FR performance. We provide pointers to a few of the recent approaches in this section.

Blanz and Vetter [2] propose a 3D morphable model-based approach in which a face is represented using a linear combination of basis exemplars. The shape and albedo parameters of the model are computed by fitting the morphable model to the input image. Romdhani et al. [3] provide an efficient and robust algorithm for fitting a 3D morphable model using shape and texture error functions. Zhang and Samaras [4] combine spherical harmonics illumination representation with 3D morphable models [2]. An iterative approach is used to compute albedo and illumination coefficients using the estimated shape. For FR across pose, local patches are considered more robust than the whole face, and several patch-based approaches have been proposed [20]. In a recent paper, Prince et al. [5] proposed a generative model for generating the observation space from the identity space using an affine mapping and pose information.

It is only recently that researchers have started looking at the problem of matching LR face images. Most of these efforts follow an SR approach. Baker and Kanade [6], [7] propose an algorithm to learn a prior on the spatial distribution of the image gradients for frontal facial images. Chakrabarti et al. [10] propose a learning-based method using kernel principal component analysis (PCA) for deriving prior knowledge about the face class for performing SR. Liu et al. [11] propose a two-step statistical modeling approach for hallucinating a HR face image from a LR input. The relationship between the HR images and their corresponding LR images is learned using a global linear model and the residual high-frequency content is modeled by a patch-based nonparametric Markov network. Xiong et al. [21] use manifold learning approaches for recovering the HR image from a single LR input. Yang et al. [15] address the problem of generating an SR image from a LR input image from the perspective of compressed sensing. A novel patch-based face hallucination framework is proposed by Tang et al. [22]. Since many FR systems use an initial dimensionality reduction method, Gunturk et al. [23] proposed eigenface-domain SR in the lower dimensional face space.

The main aim of most SR algorithms is to generate a good HR reconstruction, and they are usually not designed from a matching perspective. Recently, Hennings-Yeomans et al. [8] proposed an approach to perform SR and recognition simultaneously. Using features from the face and SR priors, they extract an HR template that simultaneously fits the SR as well as the face-feature constraints. Arandjelovic and Cipolla [24] propose a generative model for separating the illumination and down-sampling effects for the problem of matching a face in a LR query video sequence against a set of HR gallery sequences. Recently, an MDS-based approach [25] was used for improving the matching performance of LR images assuming that the probe images are in the same pose and resolution as the gallery images. Given an LR face image, Jia and Gong [26] propose directly computing a maximum likelihood identity parameter vector in the HR tensor space, which can be used for recognition and reconstruction of HR face images. There also has been some research on FR across blur [27].

3 Face Representation

Recently, there has been growing interest in using local features like SIFT [28], SURF [29], and so on, for matching facial images. Local feature descriptors describe a pixel in an image through its local neighborhood content, and recent studies have shown the effectiveness of local features for the task of FR in unconstrained environments with variations in pose and illumination [28]. Additionally, unlike most holistic face representations, these local descriptors allow for direct comparison of images across resolution with suitable scale changes while computing the descriptors. Though these features are known to be robust to changes in pose and scale, they have never been used to match LR face images with an HR gallery with considerable pose and illumination difference.

To analyze the robustness of SIFT features across significant variations of these external imaging factors, we conduct a recognition experiment on the Multi-PIE dataset [1] with HR frontal gallery and LR probe images in different poses. We use SIFT descriptors at fiducial locations (top row of Fig. 1) as the features for performing recognition. Suitable scale changes in SIFT descriptor computation are made to make the comparison across resolution feasible. The SIFT descriptors from all fiducial locations are stacked together to form one global face descriptor. Fig. 1 (bottom row) shows the recognition accuracy for LR probe images in different poses as shown in the top row when compared with HR frontal gallery. In this experiment, the LR probe images are of size 20×18 while the HR gallery images are of size 60×55 (scale factor of three). The solid red line represents the recognition accuracy when the probe images are of the same resolution and pose as the gallery images. The recognition accuracy for probe images with decreasing resolutions is shown in Fig. 2. For this experiment, the pose of the probe images is fixed at 05_0 (as labeled in Multi-PIE data).

From the above analysis, we see that there are two challenges in performing FR using such local features computed at fiducial locations:

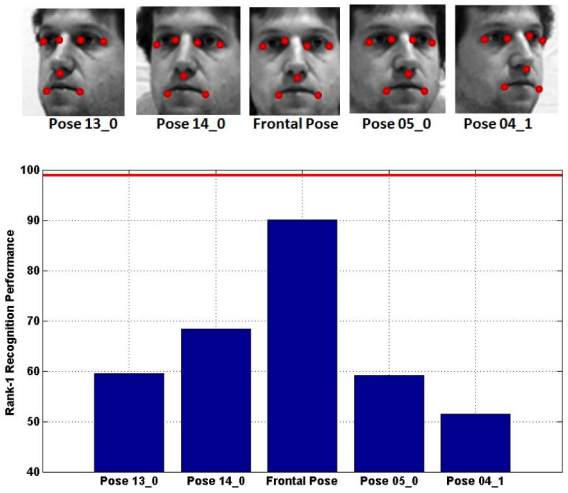

Fig. 1. (Top) Fiducial locations used for extracting SIFT descriptors for representing the input faces. (Bottom) Recognition accuracy on the Multi-PIE dataset for LR probe images with different poses as shown in the top row, but HR frontal gallery.

1. Though SIFT descriptors are fairly robust to modest variations in pose and resolution, large variations in these factors between the probe and the gallery images result in significant degradation in recognition performance.
2. Locating the facial landmarks in LR, nonfrontal images under varying illumination conditions is by itself a challenging task.

In this work, we propose a completely automatic MDS-based approach to improve the recognition performance of matching probe images with significant differences in these external factors as compared to the enrolled gallery images. Tensor analysis is used to estimate the approximate pose and the rough locations of facial landmarks.

In the following analysis, we assume that the fiducial landmarks have already been localized. In Section 5, we will discuss how to automatically estimate the approximate locations of the facial landmarks using tensor analysis.

4 Transformation Learning

Our goal is to transform SIFT descriptors extracted from the HR gallery and LR probe images to a space in which their inter-euclidean distances approximate the distances had all the descriptors been computed using HR frontal images. Note that the SIFT descriptors from all fiducial locations are stacked together to form one global face descriptor. In the proposed approach, the desired transformation is learned using MDS. Side information of feature distances between images had they all been of the same resolution and pose is provided to assist in learning the transformation.

An image is denoted by $\mathbf{I}^{(x,y)}$, where x can take values from $\{h, l\}$ to denote a HR or LR image, respectively. The variable y can take values from $\{f, p\}$ depending on whether the images are in the frontal or a nonfrontal pose. According to our notation, the HR frontal images are denoted by $\mathbf{I}_i^{(h,f)}$, $i = 1, 2, \ldots, N$, and the LR nonfrontal images are denoted by $\mathbf{I}_i^{(l,p)}$, where N is the number of images. Similarly, $\mathbf{x}_i^{(h,f)}$ and $\mathbf{x}_i^{(l,p)}$ denote the corresponding

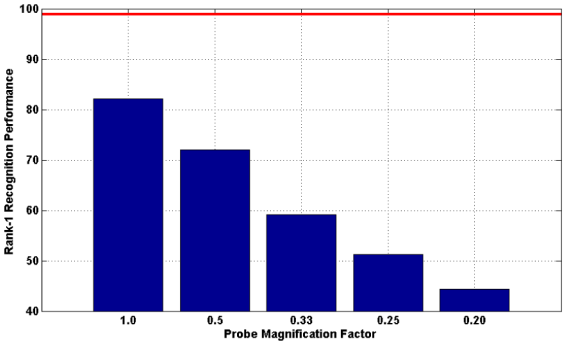

Fig. 2. Recognition accuracy for nonfrontal probe images with decreasing resolutions when compared against frontal HR gallery. Fig. 12 shows example images at these resolutions.

SIFT-based feature descriptors. Training data consisting of HR images in frontal pose are used to provide the side information required to learn the desired transformation. The distance between the features $\mathbf{x}_i^{(h,f)}$ and $\mathbf{x}_j^{(h,f)}$ from the HR frontal images is denoted by $d_{i,j}^{(h,f)}$. Clearly, the best scenario is when all the input images are of the same resolution and are in the same pose. As shown in Figs. 1 and 2, variations in the different extrinsic factors affect the recognition performance adversely. Here, we want to transform the features from the nonfrontal LR images ($\mathbf{x}_j^{(l,p)}$) and frontal HR images ($\mathbf{x}_i^{(h,f)}$) in such a way that the distances between them approximate the distances given by $d_{i,j}^{(h,f)}$, which would have been the case had the images been in the same pose and resolution. The process of learning the desired mapping is described below.

4.1 Computation of Transformation Matrix

Let $\mathbf{f} : R^d \rightarrow R^m$ denote the mapping from the input feature space R^d to the embedded euclidean space R^m. Here, m is the dimension of the transformed space and d denotes the input dimension. We consider the mapping $\mathbf{f} = (f_1, f_2, \ldots, f_m)^T$ to be a linear combination of p basis functions of the form

$$f_i(\mathbf{x}; \mathbf{W}) = \sum_{j=1}^{p} w_{ji}\phi_j(\mathbf{x}), \tag{1}$$

where $\phi_j(\mathbf{x})$, $j = 1, 2, \ldots, p$, can be a linear or nonlinear function of the input feature vectors. Here, $[\mathbf{W}]_{ij} = w_{ij}$ is the $p \times m$ matrix of the weights to be determined. The mapping defined by (1) can be written in a compact manner as follows:

$$\mathbf{f}(\mathbf{x}; \mathbf{W}) = \mathbf{W}^T\phi(\mathbf{x}). \tag{2}$$

The goal is to simultaneously transform the feature vectors from $\mathbf{I}_i^{(h,f)}$ and $\mathbf{I}_j^{(l,p)}$ such that the euclidean distance between the transformed feature vectors approximates the best possible distance $d_{i,j}^{(h,f)}$. To this end, we find the transformation $\mathbf{W}$ which minimizes the following objective function:

$$\mathbf{J}(\mathbf{W}) = \lambda\mathbf{J}_{\mathrm{DP}}(\mathbf{W}) + (1 - \lambda)\mathbf{J}_{\mathrm{CS}}(\mathbf{W}). \tag{3}$$

The first term $\mathbf{J}_{\mathrm{DP}}$ is the distance preserving term, which ensures that the distance between the transformed feature vectors approximates the distance $d_{i,j}^{(h,f)}$ and is given by

$$\mathbf{J}_{\mathrm{DP}}(\mathbf{W}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(q_{ij}(\mathbf{W}) - d_{i,j}^{(h,f)}\right)^2. \quad (4)$$

Here, $q_{ij}(\mathbf{W}) = |\mathbf{W}^T\{\phi(\mathbf{x}_i^{(h,f)}) - \phi(\mathbf{x}_j^{(l,p)})\}|$ is the distance between the transformed feature vectors of the images $\mathbf{I}_i^{(h,f)}$ and $\mathbf{I}_j^{(l,p)}$. The second term of the objective function $\mathbf{J}_{\mathrm{CS}}$ is an optional class separability term to further facilitate discriminability. Here, as in [30], we use a simple class preserving term that tries to minimize the distance between feature vectors belonging to same class and is of the form

$$\mathbf{J}_{\mathrm{CS}}(\mathbf{W}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\delta(\omega_i,\omega_j)q_{i,j}^2(\mathbf{W}), \quad (5)$$

where $\delta(\omega_i,\omega_j) = 0$ when $\omega_i \neq \omega_j$ and 1 otherwise. Here, ω_i denotes the class label of the ith image. Clearly, the distance $q_{i,j}(\mathbf{W})$ and thus the objective function depend on the transformation matrix $\mathbf{W}$. The relative effect of the two terms in the objective function is controlled by the parameter λ. Separating the terms containing $\mathbf{W}$, the final objective function takes the form

$$\mathbf{J}(\mathbf{W}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{i,j}\left(q_{i,j}(\mathbf{W}) - \beta_{i,j}d_{i,j}^{(h,f)}\right)^2. \quad (6)$$

Here, $\alpha_{i,j} = (1-\lambda)\delta(\omega_i,\omega_j) + \lambda$ and $\beta_{i,j} = \lambda/\alpha_{i,j}$. Next, we describe the algorithm which can be used for minimization of functions of this form.

4.2 Iterative Majorization Algorithm

The iterative majorization algorithm [30], [14] is used to minimize the objective function (6) to solve for the transformation matrix $\mathbf{W}$. The central idea of the majorization method is to iteratively replace the original function $\mathbf{J}(\mathbf{W})$ by an auxiliary function, also called the majorization function, which is simpler to minimize than the original function. The different steps of the majorization algorithm are given in Fig. 3. Please refer to [14] for details of the algorithm.

To perform recognition or verification, the SIFT descriptors of the gallery and probe images are transformed using the learned transformation, followed by computation of euclidean distances between the transformed features.

Input: Objective function $\mathbf{J}(\mathbf{W})$.
Output: Transformation matrix $\mathbf{W}$.

Procedure:
1. Start iteration with t = 0. Initialize $\mathbf{W}$ with random values uniformly over the range $[-1,+1]$.
2. Set $\mathbf{V} = \mathbf{W}^t$.
3. Update $\mathbf{W}^t$ to $\mathbf{W}^{t+1}$, where $\mathbf{W}^{t+1}$ is the solution that minimizes the majorization function and is given by

$$\mathbf{W} = \mathbf{A}^{-1}\mathbf{C}(\mathbf{V})\mathbf{V}$$

where $\mathbf{A}^{-1}$ is the Moore-Penrose inverse of $\mathbf{A}$. The terms $\mathbf{A}$ and $\mathbf{C}(\mathbf{V})$ are given by

$$\mathbf{A} = \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{i,j}(\phi_i-\phi_j)(\phi_i-\phi_j)^T$$

$$\mathbf{C}(\mathbf{V}) = \sum_{i}\sum_{j}c_{i,j}(\mathbf{V})(\phi_i-\phi_j)(\phi_i-\phi_j)^T$$

Here

$$c_{i,j}(\mathbf{V}) = \begin{cases} \lambda d_{i,j}^{(h,f)}/q_{i,j}(\mathbf{V}); & q_{i,j}(\mathbf{V}) > 0 \\ 0; & q_{i,j}(\mathbf{V}) = 0 \end{cases}$$

4. Check for convergence. If convergence criterion is not met, set $t = t+1$ and go to step 2, otherwise stop the iteration and output the current $\mathbf{W}$.

Fig. 3. Iterative majorization algorithm.

5 Automatic Feature Localization

In this section, we will discuss how to automatically determine the approximate feature locations in the LR uncontrolled probe images. Following the approach presented in [31], we apply multilinear analysis to the facial image data using N-mode decomposition. Suppose the image data consist of K LR images of resolution $m \times n$ (i.e., an image is represented as a $M = m \times n$ vector of image pixels) under P different poses and L different illumination conditions, then the facial image data D is of dimension $K \times P \times L \times M$. The 4-mode decomposition of $\mathcal{D}$ is given by

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{people} \times_2 \mathbf{U}_{views} \times_3 \mathbf{U}_{illum} \times_4 \mathbf{U}_{pixels}. \quad (7)$$

Here, the $K \times P \times L \times M$ core tensor $\mathcal{Z}$ governs the interaction between the different factors represented in the four mode matrices. The $K \times K$ mode matrix $\mathbf{U}_{people}$ spans the space of people parameters, the $P \times P$ mode matrix $\mathbf{U}_{views}$ spans the space of viewpoint parameters, and the $L \times L$ mode matrix $\mathbf{U}_{illum}$ spans the space of illumination parameters. The $M \times N$ $(N = K \times P \times L)$ mode matrix $\mathbf{U}_{pixels}$ orthonormally spans the space of images. Each column of $\mathbf{U}_{pixels}$ is an *eigen image*, which is identical to conventional Eigenfaces [32]. In multilinear analysis, the core tensor $\mathcal{Z}$ can transform the eigenimages in $\mathbf{U}_{pixels}$ into TensorFaces, which represent the principal axes of variation across the various modes (people, viewpoints, illuminations) and represent how the various factors interact with each other to create the face images. The $K \times P \times L \times M$ tensor $\mathcal{B}$ given by

$$\mathcal{B} = \mathcal{Z} \times_2 \mathbf{U}_{views} \times_3 \mathbf{U}_{illum} \times_4 \mathbf{U}_{pixels} \quad (8)$$

defines $P \times L$ different bases for each combination of viewpoint and illumination. Each of these bases have K eigenvectors that span the people space in which the first eigenvector depicts the average person and the remaining eigenvectors capture the variability across people for the particular combination of viewpoint and illumination. This tensor $\mathcal{B}$ is computed from LR training images. Fig. 4 shows a few facial images of a subject from the Multi-PIE data with variations in pose and illumination conditions used for doing the tensor analysis. The three tensor basis corresponding to two different illuminations for all five poses are shown in Figs. 5 (left) and 5 (right), respectively.

Given an LR probe image under unknown pose and illumination condition, we first estimate the pose using the precomputed $\mathcal{B}$. Let $\mathcal{B}_{p,l}$ denote the subtensor corresponding

Fig. 4. A few facial images of a subject from the Multi-PIE data with variations in pose and illumination conditions.

to pose p and illumination l which is flattened along the people mode to obtain the $K \times M$ matrix $\mathbf{B}_{p,l(people)}$. Now, given the vectorized probe image d, the projection operator $\mathbf{B}^{-T}_{p,l(people)}$ is used to project d into a set of candidate coefficient vectors for every combination of pose and illumination given by

$$\mathbf{c}_{p,l} = \mathbf{B}^{-T}_{p,l(people)} d. \tag{9}$$

We use the coefficient vectors to compute the reconstructed image and reconstruction error. The pose that results in the minimum reconstruction error (across all illuminations) is taken as the estimated pose of the probe image.

For a probe image, given the estimated pose, the initial fiducial locations are taken to be the median locations corresponding to the estimated pose. Fig. 6 shows the median fiducial locations for different probe poses and the distribution of the locations for the training images superimposed on one image from each pose. To account for the subtle differences in these locations for different subjects, the fiducial locations are perturbed slightly and the perturbed location which has the minimum distance to the gallery is chosen. For this, all the fiducial locations are treated separately, and thus the desired MDS-based transformations are learned for each of the locations instead of in a combined manner. The advantage of treating the locations separately is that the increase in computational cost due to perturbations of the locations is limited. A flowchart of the proposed MDS-based approach with both training and testing phases is shown in Fig. 7.

6 Experimental Evaluation

In this section, we report the results of extensive experiments performed to evaluate the usefulness of the proposed

Fig. 5. Left: The three tensor basis corresponding to illumination 7 (as labeled in the Multi-PIE data set) for the different poses. Right: Tensor basis corresponding to illumination 2.

Fig. 6. Distribution of the fiducial locations for different probe poses corresponding to the training images and the median locations.

algorithm. The experiments are divided into two parts. First, we show how the proposed approach performs with manually supervised feature localizations. We have used Active Shape Model-based (freely available) C++ software library called STASM [33] to detect feature locations automatically. The detections were eyeballed to verify that the locations were approximately correct and the incorrect locations were manually corrected. The experiments are designed to answer the following questions:

- How does the proposed approach perform when directly comparing HR frontal gallery images with LR probe images captured under uncontrolled pose and illumination conditions?
- How does the approach compare against the typical approach of performing SR on the LR probe images to allow comparison with the HR gallery?
- How does the approach compare against state-of-the-art learning-based classifiers performing the same task?
- How does the approach perform across different resolutions of the probe images?

Second, we report the performance using the completely automatic tensor analysis-based pose estimation and feature point localization. We also show the usefulness of the proposed approach for the task of simultaneous tracking and recognition of faces in surveillance videos. We observe that a learning-based measurement likelihood based on the proposed approach outperforms the traditional appearance modeling approaches on both recognition and tracking accuracy metrics.

6.1 Datasets Used and Experimental Settings

Most experiments described in this paper are performed on CMU Multi-PIE face dataset [1]. The dataset contains images of 337 subjects who attended one to four different recording sessions, which were separated by at least a month. The images were taken under different illumination conditions, pose, and expressions. For our experiments, we use images in pose 04_1, 05_0, 13_0, and 14_0 (as shown in Fig. 1 (top row)) in addition to frontal pose for gallery. Images from all 20 illuminations in the dataset with neutral expression are used for evaluation. Images of 100 randomly chosen subjects are used for training and the remaining subjects for testing. Thus, there is no subject overlap across training and test sets. The aligned face images are downsampled from the original resolution to lower resolutions using standard bicubic interpolation technique.

We also report performance of the proposed approach on the Surveillance Cameras Face Database [17]. The Multiple Biometric Grand Challenge (MBGC) [34] video challenge data is used to evaluate the performance of the proposed approach for the simultaneous tracking and recognition

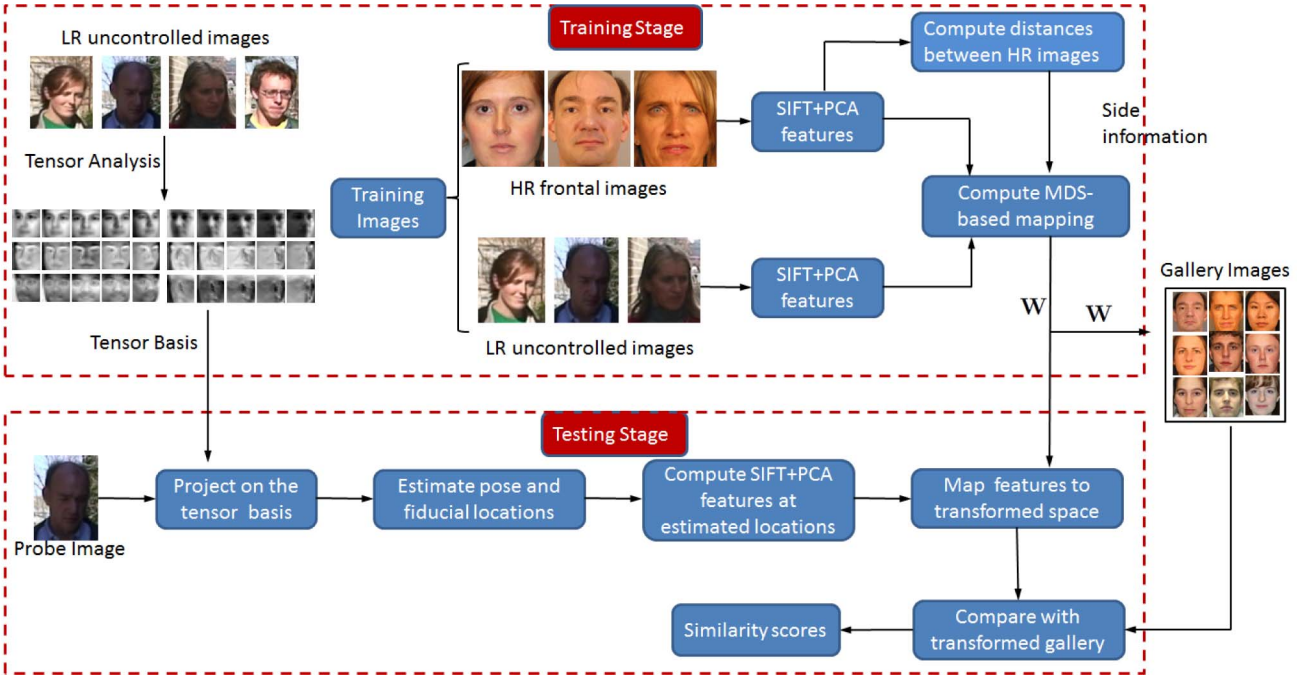

Fig. 7. A flowchart of the proposed approach.

application. Brief descriptions of both the data sets are provided later along with the details of the experiments.

Recognition experiments are conducted across illumination conditions with images from one illumination forming the gallery, while images from a different illumination forming the probe set. Thus, there is just one image per subject in the gallery and probe sets. Reported recognition accuracy is the average rank-1 recognition performance averaged over all $\binom{20}{2}$ pairs of illumination conditions forming gallery and probe sets. In this protocol, any two images compared differ in resolution, pose and illumination condition.

Similarly to [28], we use SIFT descriptors computed at the fiducial locations (as shown in Fig. 1) as the input feature. The SIFT descriptors from all fiducial locations are stacked together to form one global face descriptor. PCA is used to reduce the dimensionality of the input features, and the number of PCA coefficients is determined based on the number of eigenvalues required to capture 98 percent of the total energy. SIFT descriptors from FRGC training data [35] consisting of 366 face images are used to generate the PCA space. For all experiments, the kernel mapping ϕ is set to identity (i.e., $\phi(\mathbf{x}) = \mathbf{x}$) to highlight just the performance improvement due to the proposed learning approach. For training, the weights w_{ij} of the transformation matrix were initialized with random values uniformly over the range $[-1, +1]$. We have seen that the objective function decreases till around 20 iterations and then stabilizes. The value of the parameter λ is set to 0.5 and the output dimension m is set to 50.

6.2 Recognition across Resolution, Pose, and Illumination

First, we perform a recognition experiment with HR frontal images as the gallery and LR images in different poses and illuminations as the probe set. The resolution of the gallery images is 60×55, while that of the probe images is 20×18 (scale factor 3) for this experiment. Unless otherwise stated, the same gallery and probe resolutions are used for the other experiments. For each gallery illumination, Table 1 shows the rank-1 recognition performance averaged over the probe images under all illumination conditions. The recognition performance using SIFT+PCA directly without the proposed learning is also given as baseline for comparison. Fig. 8

TABLE 1
Rank-1 Recognition Percentage for Different Gallery Illumination Averaged over the Probe Images under All Illuminations

Gallery Illum. Probe Pose	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
13_0 (SIFT)	61	54	58	60	60	60	57	53	54	57	59	58	58	56	67	68	64	64	63	61
13_0 (Ours)	**76**	**67**	**76**	**81**	**81**	**76**	**75**	**70**	**72**	**76**	**78**	**76**	**76**	**72**	**85**	**81**	**76**	**77**	**79**	**76**
14_0 (SIFT)	71	64	68	68	70	71	68	62	63	68	68	66	64	58	76	76	73	72	71	71
14_0 (Ours)	**83**	**73**	**82**	**88**	**88**	**87**	**85**	**82**	**82**	**87**	**89**	**84**	**81**	**74**	**91**	**91**	**87**	**89**	**89**	**83**
05_0 (SIFT)	63	48	51	55	52	53	55	55	58	65	62	62	61	55	60	60	62	73	71	62
05_0 (Ours)	**81**	**75**	**81**	**84**	**82**	**80**	**81**	**79**	**80**	**83**	**82**	**83**	**79**	**70**	**85**	**84**	**83**	**85**	**86**	**80**
04_1 (SIFT)	51	51	51	50	47	46	46	45	51	57	53	54	52	46	51	52	54	61	61	50
04_1 (Ours)	**72**	**67**	**75**	**77**	**71**	**71**	**69**	**68**	**71**	**75**	**78**	**76**	**70**	**62**	**77**	**73**	**71**	**79**	**80**	**71**

Here, the HR frontal gallery is compared to LR probe images in different nonfrontal poses.

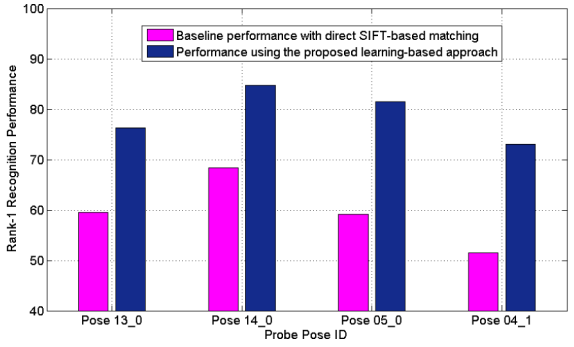

Fig. 8. Recognition performance for HR frontal gallery and LR probe images for different nonfrontal poses.

shows the recognition performance for HR frontal gallery and LR probe images in different nonfrontal poses averaged over all gallery illuminations. The proposed approach significantly improves the recognition accuracy as compared to directly using the SIFT+PCA features.

Since SIFT features are robust to illumination variations [1] and, during the training stage, images acquired under different illumination conditions are used to learn the transformation matrices, the proposed approach is robust to illumination variations. To verify this, we performed an experiment in which the pose is kept fixed at 05_0 and only the illumination varies. The performance of the proposed approach (96.1) is better as compared to the baseline (90.1) and is very close to the ideal HR-HR performance (98.9).

6.3 Comparison with SR Approach

For matching LR images, the most commonly used approach is to first obtain an HR image using SR techniques and use the super-resolved images for matching. For comparison, we use two different state-of-the-art SR techniques [15], [36] to obtain HR images from the input LR probe images. The two techniques are briefly described below:

1. *Sparse Representation-based SR* (SR1) [15]. Here, the different patches of the HR image are assumed to have a sparse representation with respect to an overcomplete dictionary of prototype signal atoms. The principle of compressed sensing is used to correctly recover the sparse representation from the down-sampled input image. For our experiments, we have used the code and the pretrained dictionary available from the author's website [37].
2. *Regression-based method* (SR2) [36]. Here, the basic idea is to learn a mapping from input LR images to target HR images from example image pairs using kernel ridge regression. To remove the blurring and ringing effects around strong edges because of the regression, a natural image prior that takes into account the discontinuity property of images is used for postprocessing. Code available on the author's website is used for this technique [38].

Given the LR nonfrontal probe images, the SR approach is used to compute HR images that are then used for computing the SIFT features. Fig. 9b shows examples of LR probe images, while Figs. 9c and 9d show the corresponding outputs of SR1 and SR2 algorithms. The original HR images are shown in Fig. 9a. Fig. 10 shows the recognition performance obtained. Though the SR algorithms improve upon the baseline performance, we observe that the proposed approach performs considerably better than the state-of-the-art SR approaches. The performance improvement is even more significant when we use the features computed from the SR images in the proposed MDS-based approach, implying that SR algorithms can be used along with the proposed approach for further performance improvement.

Fig. 9. (a) Original HR images. (b) Input LR images. (c), (d) Output images of the SR1 [15] and SR2 [36] algorithms with scale factor 3.

6.4 Comparison with Classifier-Based Approach

Recently, metric learning approaches like Large Margin Nearest Neighbor (LMNN) [16] on local features like SIFT have been used successfully for recognizing faces in unconstrained environments [28]. LMNN is used to learn a Mahalanobis distance metric for k-nearest neighbor (kNN) classification by semi-definite programming. Fig. 11 compares the performance of the proposed approach with LMNN for HR frontal gallery and LR probes in two different poses. We use the code available from author's website [39]. With default settings, the performance of LMNN was worse than the baseline. We experimented with different settings of validation and number of nearest

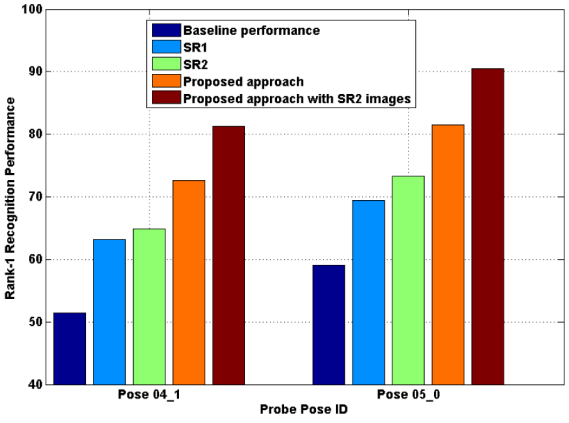

Fig. 10. Comparison with two SR approaches for two different probe poses. The bars are in the order of the legend.

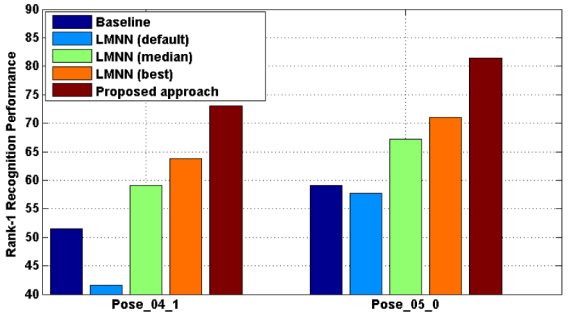

Fig. 11. Comparison of the proposed approach with LMNN [16] for HR gallery and LR probe images for two different poses. The bars are in the order of the legend.

neighbor parameters and report the median and best performance obtained. We see that for large variations of scale and pose between the gallery and probe images, the proposed approach performs considerably better than this state-of-the art learning approach.

6.5 Performance with Different Probe Resolutions

In all the experiments so far, the scale factor between the HR gallery and LR probe images was fixed at 3. Here, we analyze the performance of the proposed approach for varying resolutions of the probe images. The gallery resolution is fixed at 60×55 and four different probe resolutions are considered, 32×28, 21×19, 16×14, and 13×11. LR probe images in pose 05_0 are used for this experiment. Fig. 12 shows example gallery and probe images. The bottom row of Fig. 12 shows the recognition performance using the SIFT+PCA features and the proposed approach. We see that the proposed approach is successful in significantly improving the recognition performance for a wide range of resolutions.

6.6 Performance with Automatically Localized Fiducial Locations

All the experiments conducted so far were on manually supervised fiducial locations. Now, we present the results

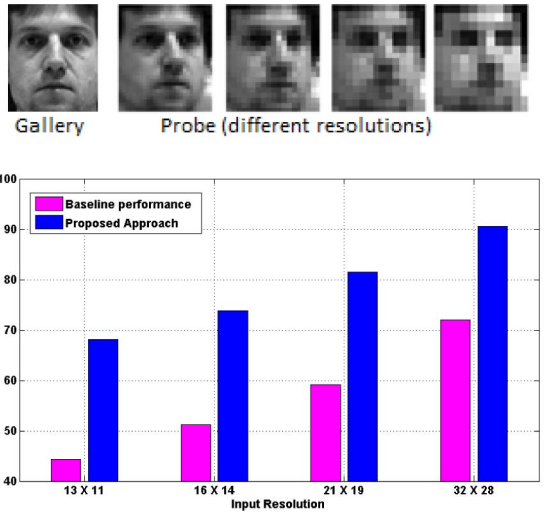

Fig. 12. (Top) Gallery and probe images at different resolutions. (Bottom) Recognition performance of the baseline and the proposed approach for different probe resolutions.

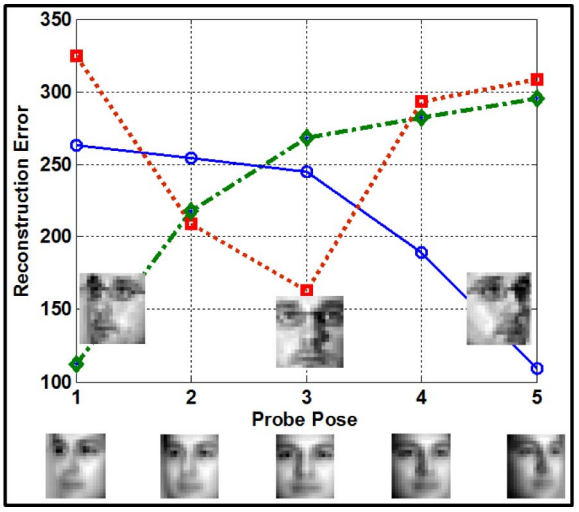

Fig. 13. Pose reconstruction error for probe images in different head poses. The pose resulting in the least reconstruction error is taken to be the estimated pose of the probe image. The plot shows error curves for three different probe images, which are placed adjacent to the corresponding curves. The bottom row shows the average faces in various poses.

using automatically computed fiducial locations obtained as discussed in Section 5.

First, tensor analysis is used to estimate the head pose of a given probe image. The pose with the lowest reconstruction error is taken as the estimated pose. Fig. 13 shows pose reconstruction error curves for probe images in different poses. Table 2 shows the accuracy of this pose estimation technique for images from the Multi-PIE dataset. In this experiment, the tensor is formed from randomly chosen 100 subjects, and images of the remaining subjects under the five different poses and 20 different illumination conditions are used as the probe images (i.e., a total of $237 \times 5 \times 20 = 23{,}700$ images). We see that this technique for estimating the approximate head pose works quite well with an average percentage error of 0.73 percent for all the poses, i.e., only 0.73 percent of all the probe images were classified incorrectly to a wrong pose.

For each probe image, given the estimated pose, the initial fiducial locations are taken to be the median locations (Fig. 6) corresponding to the estimated pose. The fiducial locations are perturbed slightly to account for the subtle differences in these locations for different subjects and the perturbed location which has the minimum distance to the gallery images is chosen. The matching results shown so far stack SIFT descriptors from all fiducial locations and use that as a global face descriptor to perform the proposed MDS-based transformation learning. Performing location perturbation in such a setting will lead to combinatorial increase in computational cost. To this end, we use a variant in which all the fiducial locations are treated separately, and

TABLE 2
Average Error in Pose Estimation on the Multi-PIE Data Using Tensor Analysis

Pose 1	Pose 2	Pose 3	Pose 4	Pose 5	Average
0.85%	0.85%	0.80%	0.70%	0.45%	0.73%

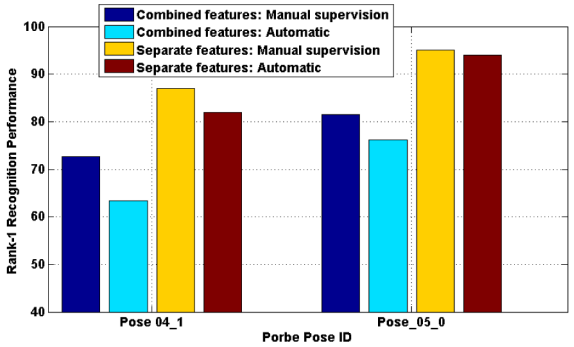

Fig. 14. Comparison of recognition performance using manually supervised fiducial point localizations with the completely automatic approach using tensor analysis. The bars are in the order of the legend.

thus the transformations are learned for each of the locations instead of in a combined manner as done before. This facilitates location perturbation to be applied to account for small errors in estimated fiducial locations without prohibitive increase in matching cost.

Fig. 14 shows the recognition performance obtained using the automatically estimated fiducial locations using the tensor analysis. Rank-1 recognition rates obtained using both stacked SIFT representation (no perturbation) and separate SIFT representation are shown for comparison. Performance obtained using the fiducial locations obtained using STASM library followed by manual supervision are also included. Following observations can be made from this comparison. As expected, using automatically estimated fiducial locations results in slightly worse performance, but the performance is still significantly better than the baseline approach of directly using SIFT+PCA feature. The degradation in performance is worse using stacked SIFT representation. This can probably be attributed to the inability to perform small perturbations in such a representation to account for inaccuracies in estimated fiducial locations. When the feature locations are considered separately (along with perturbations for the automatic case), improved performance is obtained for both the manually supervised and automatic fiducial locations.

6.7 Evaluation on Surveillance Quality Images

We now test the usefulness of the proposed approach on the Surveillance Cameras Face Database [17], which contains images of 130 subjects taken in an uncontrolled indoor environment using five video surveillance cameras of various qualities. As in typical commercial surveillance systems, the database was collected with the camera placed slightly above the subject's head, and also the individuals were not required to look at a fixed point during the recordings, thus making the data even more challenging. The gallery images were captured using a high-quality photo camera. We use images from all five surveillance cameras, resulting in a total of 650 probe images. Fig. 15 shows sample gallery (top row) and probe images (bottom row) of a few subjects. Most of the probe images in the dataset have more extreme poses than the ones in these examples, but those images could not be released according to the Database Release Agreement.

For the experiment, we use images of randomly picked 50 subjects for training and the remaining 80 subjects for testing. Thus, there is no identity overlap between the

Fig. 15. First Row: Example gallery images. Second Row: Example probe images of the same subjects as the gallery images [17]. Most of the probe images in the dataset have more extreme poses than the ones in these examples, but those images could not be released according to the Database Release Agreement.

training and test sets. The experiment is repeated 10 times with different random sampling of the subjects. Fig. 16 shows the Cumulative Match Characteristic (CMC) curves for this experiment using SIFT+PCA coefficients as the input feature. The error bars indicate the variation in performance for different runs of the experiment. The number of PCA coefficients is determined based on the number of eigenvalues required to capture 98 percent of the total energy. Performance using other popular local features like SURF [29] and LBP codes [40] is also shown for comparison. We see that the proposed approach significantly outperforms the SIFT+PCA baseline and all other compared approaches in matching real surveillance images. The proposed approach also compares favorably against SIFT+LDA, as shown in the figure.

7 Application to Tracking and Recognition in Surveillance Videos

For video-based FR, a tracking-then-recognition paradigm is typically followed in which the faces are first tracked and

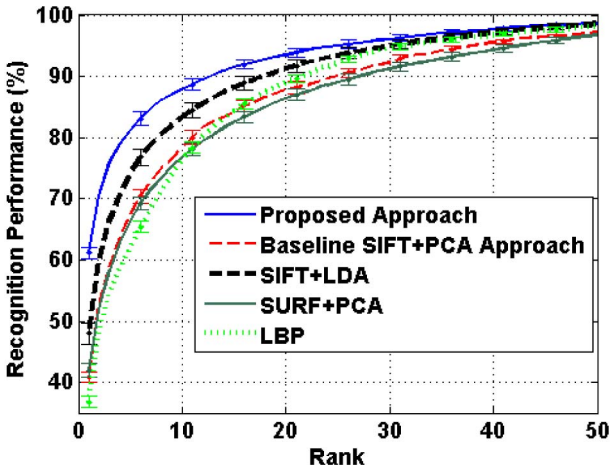

Fig. 16. CMC curves obtained using the proposed approach and the baseline approach on Surveillance Cameras Face Database [17]. Comparisons with SIFT+LDA [41] and other local feature-based approaches like SURF [29] and LBP [40] are also shown. The experiment is repeated 10 times with different random sampling of the subjects to form training and test sets (with no identity overlap). The error bars indicate the variation in performance observed for different runs of the experiment.

Fig. 17. A few example frames from a video from the MBGC video challenge [34].

then used for recognition. But both tracking and recognition are very challenging for low-quality surveillance videos with LR and significant variations in pose and illumination. Here, we extend the simultaneous tracking and recognition framework [42], which performs the two tasks of tracking and recognition in a single unified framework to address these challenges. We show that a learning-based likelihood measurement model based on the proposed approach can be used to improve both the tracking as well as recognition performance for surveillance videos. Here, we consider the scenario in which the gallery consists of one or more HR frontal images, while the probe consists of LR videos with uncontrolled pose and illumination as is typically obtained in surveillance systems.

7.1 Simultaneous Tracking and Recognition

For completion, we briefly describe the tracking and recognition framework [42], which uses a modified version of the CONDENSATION algorithm for tracking the facial features across the frames in the poor quality probe video and for recognition. The filtering framework consists of a motion model which characterizes the motion of the subject in the video. The overall state vector of this unified tracking and recognition framework consists of an identity variable in addition to the usual motion parameters, and the observation model determines the measurement likelihood.

Motion Model. The *motion model* is given by the first-order Markov chain:

$$\theta_t = \theta_{t-1} + u_t; t \geq 1. \tag{10}$$

Here, affine motion parameters are used and so $\theta = (a_1, a_2, a_3, a_4, t_x, t_y)$, where $\{a_1, a_2, a_3, a_4\}$ are deformation parameters and $\{t_x, t_y\}$ are 2D translation parameters. u_t is noise in the motion model.

Identity equation. Assuming that the identity does not change as time proceeds, the *identity equation* is given by

$$n_t = n_{t-1}; t \geq 1. \tag{11}$$

Observation Model. Assuming that the transformed observation is a noise-corrupted version of some still template in the gallery, the *observation equation* can be written as

$$T_{\theta_t}\{z_t\} = I_{n_t} + v_t; t \geq 1, \tag{12}$$

where v_t is the observation noise at time t and $T_{\theta_t}\{z_t\}$ is a transformed version of the observation z_t. Here, $T_{\theta_t}\{z_t\}$ is composed of: 1) an affine transform of z using $\{a_1, a_2, a_3, a_4\}$, 2) cropping the region of interest at position $\{t_x, t_y\}$, and 3) performing zero-mean-unit-variance normalization.

Usually, likelihood measurement models like a truncated Laplacian or probabilistic subspace density approach are used to handle the appearance difference between the probe and the gallery [42]. Here, we propose extending the MDS-based approach for computing the measurement likelihood, which results in better modeling of the appearance difference between the gallery and probe resulting in both better tracking and recognition for surveillance videos.

To compute the measurement likelihood, the SIFT descriptors of the gallery and affine-transformed probe frame are mapped using the learned transformation $\mathbf{W}$, followed by computation of euclidean distances between the transformed features:

$$p(z_t \mid n_t, \theta_t) = \left|\mathbf{W}^T\left[\phi(T_{\theta_t}\{z_t\}) - \phi(x_{n_t})\right]\right|. \tag{13}$$

7.2 Experimental Evaluation

For our experiments, we use 50 surveillance quality videos (each 40-100 frames from 50 subjects) from the MBGC [34] video challenge data for the probe videos. Fig. 17 shows a few sample frames from a video sequence. Since the MBGC video challenge data do not contain the HR frontal still images needed to form the HR gallery set, we select images of the same subjects from FRGC data [35], which has considerable subject overlap with the MBGC data. Fig. 18 (top row) shows a few sample gallery images from the dataset used and the bottom row shows cropped face regions from the corresponding probe videos. We see that there is a considerable difference in pose, illumination, and resolution between the gallery images and the probe videos.

We compare the proposed learning-based likelihood measurement model with the following two approaches for computing the likelihood measurement [42]:

1. *Truncated Laplacian likelihood*. Here, the likelihood measurement model is given by [42]

$$p(z_t \mid n_t, \theta_t) = \mathrm{LAP}\left(\|T_{\theta_t}\{z_t\} - I_{n_t}\|; \sigma_1, \tau_1\right). \tag{14}$$

Fig. 18. (Top) Example HR gallery images. (Bottom) Cropped facial regions from the corresponding LR probe videos.

TABLE 3
Rank-1 Recognition Accuracy and Tracking Error (Pixels/Frame) Obtained Using the Proposed Approach

Method	Laplacian Likelihood	IPS Likelihood	Proposed Approach
Recognition Accuracy	24%	40%	68%
Tracking Error (pixels)	4.8	5.8	2.8

Comparison with other likelihood models is also reported.

Here, $\| \cdot \|$ is the absolute distance and

$$\mathrm{LAP}(x;\sigma;\tau) = \begin{cases} \sigma^{-1}\exp(-x/\sigma) & \text{if } x \leq \tau\sigma, \\ \sigma^{-1}\exp(-\tau) & \text{otherwise.} \end{cases}$$

2. *Probabilistic subspace density-based likelihood*. The probabilistic subspace density-based approach [43] has also been used to handle significant appearance differences between the gallery and probe [42]. The gallery and one video frame were used for constructing the intrapersonal space (IPS). Here, the measurement likelihood is given by

$$p(z_t \mid n_t, \theta_t) = \mathrm{PS}\big(T_{\theta_t}\{z_t\} - I_{n_t}\big), \tag{15}$$

$$\text{where} \qquad \mathrm{PS}(x) = \frac{\exp\big(-1/2\sum_{i=1}^{s}\big(y_i^2/\lambda_i\big)\big)}{(2\pi)^{s/2}\prod_{i=1}^{s}\lambda_i^{1/2}}.$$

Here, $\{\lambda_i, e_i\}_{i=1}^{s}$ are the top s eigenvalues and eigenvectors obtained by performing PCA [32] on IPS and $y_i = e_i^T x$ is the ith principal component of x.

We build upon the code provided in the authors website [44]. Training is performed using images from a separate set of 50 subjects. The number of particles for the particle filtering framework is taken to be 200.

The recognition performance of the proposed approach is shown in Table 3. Comparison with the two different likelihood models is also shown. Each test face video is classified as belonging to a subject in the gallery. The shown performance indicates the percentage of videos in which correct identity of the face in the video is determined. We see that the recognition performance of the proposed learning-based simultaneous tracking and recognition framework is considerably better than the other approaches due to better modeling of the appearance difference between the gallery and the probe images. To compute the tracking error, we manually marked three fiducial locations (the center of the two eyes and the bottom of the nose) of every fifth frame of each video. For each probe video, we measure the difference between the manually marked ground truth locations and the locations given by the tracker, and the tracking error is given by the average error in the fiducial locations. The mean tracking errors (in pixels) over all the probe videos for all the approaches are shown in Table 3. Fig. 19 shows the tracking results for a few frames of a probe video obtained using the proposed approach. Fig. 20 shows the average tracking error using the

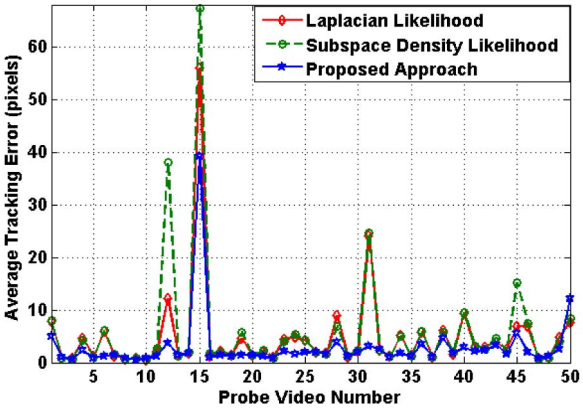

Fig. 20. Average tracking error obtained using the proposed learning-based approach. Comparison with the two other likelihood models is also shown.

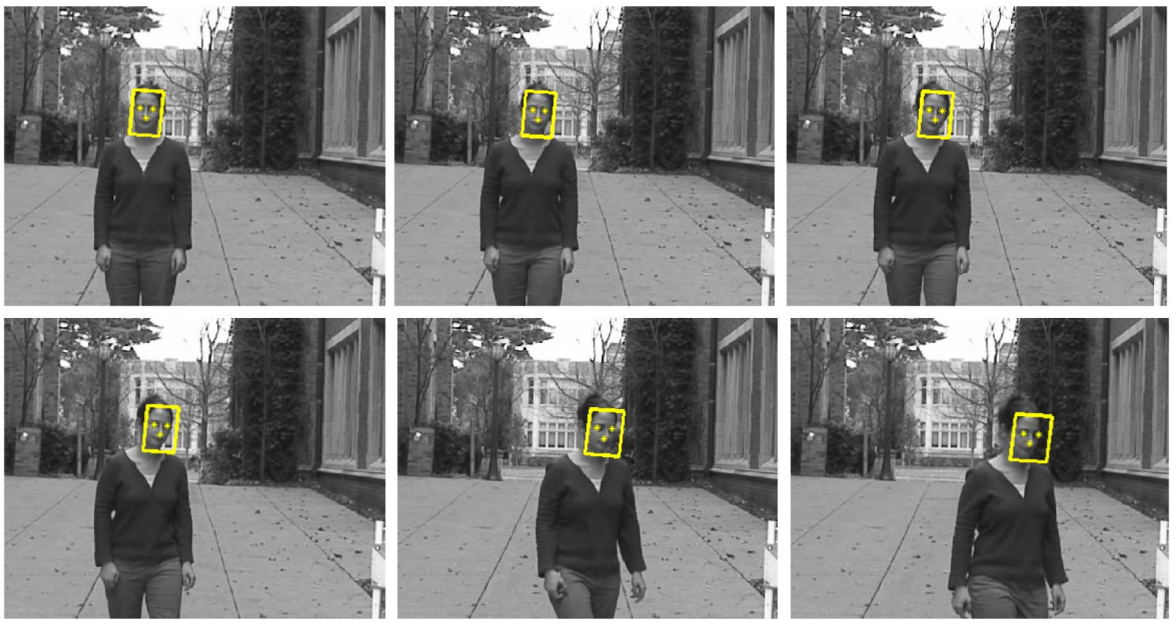

Fig. 19. A few frames showing the tracking results obtained using the proposed approach. Here, only the region of the frames containing the person is shown for better visualization.

three approaches on all 50 test videos. We see for 49 out of 50 videos that the proposed approach achieves a lower tracking error as compared to the other approaches.

8 Summary and Discussion

In this paper, we proposed a novel MDS-based approach for matching LR facial images captured from surveillance cameras with considerable variations in pose and illumination to HR gallery images in frontal pose. The basic intuition is to simultaneously transform the features from the probe and the gallery images such that the distances between them approximate the distances had the probe image been taken in the same conditions as the gallery images. Extensive evaluation on the Multi-PIE data and Surveillance Cameras Face Database shows the usefulness of the proposed approach. The application of the approach for the task of simultaneous tracking and recognition of faces in poor quality videos further signifies the practical applicability of the approach. In all the conducted experiments, the transformed features significantly outperform the baseline input features. The approach is also shown to outperform the state-of-the-art SR and classifier-based techniques. A novel tensor analysis-based approach to perform automatic pose estimation and fiducial landmark localization is also proposed.

Acknowledgments

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of IARPA, the ODNI, the Army Research Laboratory, or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Guide to the CMU Multi-Pie Database," technical report, Carnegie Mellon Univ., 2007.

[2] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 9, pp. 1063-1074, Sept. 2003.

[3] S. Romdhani, V. Blanz, and T. Vetter, "Face Identification by Fitting a 3D Morphable Model Using Linear Shape and Texture Error Functions," *Proc. European Conf. Computer Vision,* pp. 3-19, 2002.

[4] L. Zhang and D. Samaras, "Face Recognition from a Single Training Image under Arbitrary Unknown Lighting Using Spherical Harmonics," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 3, pp. 351-363, Mar. 2006.

[5] S. Prince, J. Warrell, J. Elder, and F. Felisberti, "Tied Factor Analysis for Face Recognition across Large Pose Differences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 6, pp. 970-984, June 2008.

[6] S. Baker and T. Kanade, "Hallucinating Faces," *Proc. Fourth IEEE Int'l Conf. Automatic Face and Gesture Recognition,* Mar. 2000.

[7] S. Baker and T. Kanade, "Limits on Super-Resolution and How to Break Them," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 9, pp. 1167-1183, Sept. 2002.

[8] P. Hennings-Yeomans, S. Baker, and B. Kumar, "Simultaneous Super-Resolution and Feature Extraction for Recognition of Low-Resolution Faces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[9] P. Hennings-Yeomans, B. Kumar, and S. Baker, "Robust Low-Resolution Face Identification and Verification Using High-Resolution Features," *Proc. Int'l Conf. Image Processing,* pp. 33-36, 2009.

[10] A. Chakrabarti, A. Rajagopalan, and R. Chellappa, "Super-Resolution of Face Images Using Kernel PCA-Based Prior," *IEEE Trans. Multimedia,* vol. 9, no. 4, pp. 888-892, June 2007.

[11] C. Liu, H.Y. Shum, and W.T. Freeman, "Face Hallucination: Theory and Practice," *Int'l J. Computer Vision,* vol. 75, no. 1, pp. 115-134, 2007.

[12] T. Marciniak, A. Dabrowski, A. Chmielewska, and R. Weychan, "Face Recognition from Low Resolution Images," *Proc. Int'l Conf. Multimedia Comm., Services, and Security,* pp. 220-229, 2012.

[13] W. Hwang, X. Huang, K. Noh, and J. Kim, "Face Recognition System Using Extended Curvature Gabor Classifier Bunch for Low-Resolution Face Image," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops,* pp. 15-22, 2011.

[14] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications,* second ed. Springer, 2005.

[15] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image Super-Resolution via Sparse Representation," *IEEE Trans. Image Processing,* vol. 19, no. 11, pp. 2861-2873, Nov. 2010.

[16] K.Q. Weinberger and L.K. Saul, "Fast Solvers and Efficient Implementations for Distance Metric Learning," *Proc. Int'l Conf. Machine Learning,* vol. 307, pp. 1160-1167, 2008.

[17] M. Grgic, K. Delac, and S. Grgic, "SCface—Surveillance Cameras Face Database," *Multimedia Tools and Applications J.,* vol. 51, pp. 863-879, 2009.

[18] S. Biswas, G. Aggarwal, and P. Flynn, "Pose-Robust Recognition of Low-Resolution Face Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2011.

[19] S. Biswas, G. Aggarwal, and P.J. Flynn, "Face Recognition in Low-Resolution Videos Using Learning-Based Likelihood Measurement Model," *Proc. Int'l Joint Conf. Biometrics,* 2011.

[20] T. Kanade and A. Yamada, "Multi-Subregion Based Probabilistic Approach toward Pose-Invariant Face Recognition," *Proc. IEEE Int'l Symp. Computational Intelligence in Robotics and Automation,* pp. 954-959, 2003.

[21] H. Chang, D. Yeung, and Y. Xiong, "Super-Resolution through Neighbor Embedding," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 275-282, 2004.

[22] W. Liu, D. Lin, and X. Tang, "Hallucinating Faces: Tensorpatch Super-Resolution and Coupled Residue Compensation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 478-484, 2005,

[23] B. Gunturk, A. Batur, Y. Altunbasak, M. Hayes, and R. Mersereau, "Eigenface-Domain Super-Resolution for Face Recognition," *IEEE Trans. Image Processing,* vol. 12, no. 5, pp. 597-606, May 2003.

[24] O. Arandjelovic and R. Cipolla, "A Manifold Approach to Face Recognition from Low Quality Video across Illumination and Pose Using Implicit Super-Resolution," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.

[25] S. Biswas, K.W. Bowyer, and P.J. Flynn, "Multidimensional Scaling for Matching Low-Resolution Facial Images," *Proc. IEEE Int'l Conf. Biometrics: Theory, Applications, and Systems,* 2010.

[26] K. Jia and S. Gong, "Multi-Modal Tensor Face for Simultaneous Super-Resolution and Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1683-1690, 2005.

[27] M. Nishiyama, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi, "Facial Deblur Inference to Improve Recognition of Blurred Faces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1115-1122, 2009.

[28] M. Guillaumin, J. Verbeek, and C. Schmid, "Is That You? Metric Learning Approaches for Face Identification," *Proc. IEEE Int'l Conf. Computer Vision,* 2009.

[29] P. Dreuw, P. Steingrube, H. Hanselmann, and N. Hermann, "SurfFace: Face Recognition under Viewpoint Consistency Constraints," *Proc. British Machine Vision Conf.,* 2009.

[30] A. Webb, "Multidimensional Scaling by Iterative Majorization Using Radial Basis Functions," *Pattern Recognition,* vol. 28, no. 5, pp. 753-759, May 1995.

[31] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear Analysis of Image Ensembles: Tensorfaces," *Proc. European Conf. Computer Vision,* pp. 447-460, 2002.
[32] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neurosicence,* vol. 3, no. 1, pp. 71-86, 1991.
[33] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model," *Proc. European Conf. Computer Vision,* http://www.milbo.users.sonic.net/stasm, 2008.
[34] P.J. Phillips, P.J. Flynn, J.R. Beveridge, W.T. Scruggs, A.J. O'Toole, D.S. Bolme, K.W. Bowyer, A. Draper Bruce, G.H. Givens, Y.M. Lui, H. Sahibzada, J.A. Scallan, and S. Weimer, "Overview of the Multiple Biometrics Grand Challenge," *Proc. Int'l Conf. Biometrics,* pp. 705-714, 2009.
[35] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 947-954, 2005.
[36] I.K. Kim and Y. Kwon, "Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 6, pp. 1127-1133, June 2010.
[37] http://www.ifp.illinois.edu/~jyang29/, 2013.
[38] http://www.mpi-inf.mpg.de/~kkim/, 2013.
[39] http://www.cse.wustl.edu/~kilian/code/code.html, 2013.
[40] T. Ahonen, A. Hadid, and M. Pietikinen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 2037-2041, Dec. 2006.
[41] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.
[42] S.K. Zhao, V. Krueger, and R. Chellappa, "Probabilistic Recognition of Human Faces from Video," *Computer Vision and Image Understanding,* vol. 91, pp. 214-245, 2003.
[43] B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 6, pp. 780-788, June 2002.
[44] https://sites.google.com/site/skevinzhou/codes/, 2013.

Soma Biswas received the BE degree in electrical engineering from Jadavpur University, Kolkata, India, in 2001, the MTech degree from the Indian Institute of Technology, Kanpur, in 2004, and the PhD degree in electrical and computer engineering from the University of Maryland, College Park, in 2009. She is currently working as a research assistant professor at the University of Notre Dame. Her research interests include signal, image, and video processing, computer vision, and pattern recognition. She is a member of the IEEE.

Gaurav Aggarwal received the BTech degree in computer science and engineering from the Indian Institute of Technology, Madras, in 2002, and the master's and PhD degrees in computer science from the University of Maryland, College Park, in 2004 and 2008, respectively. He is currently working as a research assistant professor at the University of Notre Dame. His research interests include image and video processing, computer vision, and pattern recognition. He is a member of the IEEE.

Patrick J. Flynn received the PhD degree in computer science in 1990 from Michigan State University. He is a professor of computer science and engineering and a concurrent professor of electrical engineering at the University of Notre Dame. He has held faculty positions at Notre Dame (1990-1991, 2001-present), Washington State University (1991-1998), and Ohio State University (1998-2001). His research interests include computer vision, biometrics, and image processing. He has served on the editorial boards of the *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Pattern Recognition*, and *Pattern Recognition Letters*. He has served as the vice-president for Finance and the vice-president for Conferences of the IEEE Biometrics Council, and was a general chair of IEEE BTAS 2012. He is a fellow of the IEEE, a fellow of the IAPR, an ACM distinguished scientist, and an associate member of the American Academy of Forensic Sciences and the International Association for Identification.

Kevin W. Bowyer received the PhD degree in computer science from Duke University. He currently serves as a Schubmehl-Prein Professor and the chair of the Department of Computer Science and Engineering at the University of Notre Dame. He previously served as editor-in-chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and currently serves as editor-in-chief of the *IEEE Biometrics Compendium*, the first IEEE virtual journal. He is the founding general chair of the IEEE International Conference on Biometrics Theory, Applications, and Systems (BTAS) conference series, served as a program chair for the 2011 IEEE International Conference on Automated Face and Gesture Recognition, and as a general chair for the 2011 International Joint Conference on Biometrics. His recent research interests include problems in biometrics and in data mining. Particular contributions in biometrics include algorithms for improved accuracy in iris biometrics, studies of basic phenomena in iris biometrics, studies involving identical twins, face recognition using three-dimensional shape, 2D and 3D ear biometrics, advances in multimodal biometrics, and support of the government's Face Recognition Grand Challenge, Iris Challenge Evaluation, Face Recognition Vendor Test 2006, and Multiple Biometric Grand Challenge programs. His paper "Face Recognition Technology: Security Versus Privacy," published in *IEEE Technology and Society*, was recognized with an "Award of Excellence" from the Society for Technical Communication in 2005. He is a fellow of the IEEE and a Golden Core member of the IEEE Computer Society.