Wide Baseline Image Registration With Application to 3-D Face Modeling

Amit K. Roy-Chowdhury, Rama Chellappa, *Fellow, IEEE*, and Trish Keaton

***Abstract*—Establishing correspondence between features in two images of the same scene taken from different viewing angles is a challenging problem in image processing and computer vision. However, its solution is an important step in many applications like wide baseline stereo, three-dimensional (3-D) model alignment, creation of panoramic views, etc. In this paper, we propose a technique for registration of two images of a face obtained from different viewing angles. We show that prior information about the general characteristics of a face obtained from video sequences of different faces can be used to design a robust correspondence algorithm. The method works by matching two-dimensional (2-D) shapes of the different features of the face (e.g., eyes, nose etc.). A doubly stochastic matrix, representing the probability of match between the features, is derived using the Sinkhorn normalization procedure. The final correspondence is obtained by minimizing the probability of error of a match between the entire constellation of features in the two sets, thus taking into account the global spatial configuration of the features. The method is applied for creating holistic 3-D models of a face from partial representations. Although this paper focuses primarily on faces, the algorithm can also be used for other objects with small modifications.**

***Index Terms*— Biometrics, face modeling, feature correspondence, image registration.**

I. Introduction

ESTABLISHING correspondence between features in two images of the same scene taken from different viewing angles is a challenging problem in image processing and computer vision. The difficulty of the problem is compounded by the fact that the images may be obtained under different conditions of lighting and camera settings. However, its solution is an important step in many applications like wide baseline stereo, three-dimensional (3-D) model alignment, creation of panoramic views, etc. Numerous methods have been tried to solve this problem, ranging from techniques which take advantage of the knowledge of the geometry of the scene to ones which use different information theoretic measures to compute similarity.

A. Literature Review

One of the well-known methods for registration is the iterative closest point (ICP) algorithm [1] of Besl and McKay. It uses a mean-square distance metric which converges monotonically to the nearest local minimum. It was used for registering 3-D shapes by considering the full six degrees of freedom in the motion parameters. It has been extended to include the Levenberg–Marquardt nonlinear optimization and robust estimation techniques to minimize the registration error [2]. Another well-known method for registering 3-D shapes is the work of Vemuri and Aggarwal where they used range and intensity data for reconstructing complete 3-D models from partial ones [3]. Registering range data for the purpose of building surface models of 3-D objects was also the focus of the work in [4]. Matching image tokens across triplets, rather than pairs, of images has also been considered. In [5], the authors developed a robust estimator for the trifocal tensor based upon corresponding tokens across an image triplet. This was then used to recover 3-D structure. Reconstructing 3-D structure was also considered in [6] using stereo image pairs from an uncalibrated video sequence. However, most of these algorithms work given good initial conditions, e.g., for 3-D model alignment, the partial models have to be brought into approximate positions. The problem of automatic "crude" registration (in order to obtain good initial conditions) was addressed in [7], where the authors used bitangent curve pairs which could be found and matched efficiently.

In the above methods, geometric properties are used to align 3-D shapes. Another important area of interest for registration schemes is two-dimensional (2-D) image matching, which can be used for applications like image mosaicing, retrieval from a database, medical imaging etc. Two-dimensional matching methods rely on extracting features or *interest points*. In [8], the authors show that interest points are stable under different geometric transformations and define their quality based on repeatability rate and information content. One of the most widely used schemes for tracking feature points is the KLT tracker [9], which combines feature selection and tracking across a sequence of images by minimizing the sum of squared intensity differences over windows in two frames. A probabilistic technique for feature matching in a multiresolution Bayesian framework was developed in [10] and used in uncalibrated image mosaicing. In [11], the authors introduced the use of Zernike orthogonal polynomials to compute the relative rigid transformations between images. It allows the recovery of rotational and scaling parameters without the need for extensive correlation and search algorithms. Precise

Manuscript received April 15, 2002; revised September 21, 2002. This work was supported in part by the National Science Foundation under Grant 0086075. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chalapathy Neti.

A. K. Roy-Chowdhury was with the Center for Automation Research University of Maryland, College Park, MD 20742 USA. He is now with the Department of Electrical Engineering, University of California, Riverside, CA 92521 USA (e-mail: amitrc@ee.ucr.edu).

R. Chellappa is with the Department of Electrical and Computer Engineering and the Center for Automation Research, University of Maryland, College Park, MD 20742 USA (e-mail: rama@cfar.umd.edu).

T. Keaton is with the Department of Signal and Image Processing HRL Laboratories LLC, Malibu, CA 90265 USA (e-mail: pakeaton@hrl.com).

Digital Object Identifier 10.1109/TMM.2004.827511

TABLE I
DESCRIPTION OF THE TEST DATABASE AND THE MEASURE OF QUALITY OF THE CORRESPONDENCE MATRIX

Subject Index	Data Quality	KL Divergence	Subject Index	Data Quality	KL Divergence
1	Training, Fig. 1(a)	6.47	2	Training, Fig. 1(b)	6.45
3	Training, Fig. 1(c)	6.77	4	Fig. 1(d)	9.58
5	Fig. 1(e)	9.49	6	Fig. 1(f)	9.53
7(1a)	Good Data	10.23	8(1b)	Glasses	11.23
9(1c)	Good Data	9.62	10(1d)	Glasses	11.43
11(1e)	Poor Lighting	10.74	12(1f)	Poor Lighting	10.62
13(1h)	Female, Glasses	12.24	14(1i)	Female	10.6
15(1j)	Glasses	11.28	16(1k)	Eyes hidden by hair	11.05
17(1l)	Glasses	11.38	18(1m)	Good Data	10.42
19(1n)	Good Data	9.99	20(1p)	Glasses and Beard	12.35
21(1q)	Glasses and Beard	12.53	22(1r)	Good Data	10.34
23(1s)	Glasses	11.24	24(1t)	Female	11.12

registration algorithms are required for medical imaging applications also. A mutual information criterion, optimized using the simulated annealing technique, was used in [12] for aligning images of the retina.

Various probabilistic schemes have also been used for solving registration problems. One of the most well-known techniques is the work of Viola and Wells for aligning 2-D and 3-D objects by maximizing mutual information [13]. The technique is robust with respect to the surface properties of objects and illumination changes. A stochastic optimization procedure was proposed for maximizing the mutual information. A probabilistic technique for matching the spatial arrangement of features using shape statistics was proposed in [14]. Most of these techniques in image registration work for rigid objects. The constraints using intensity and shape usually break down for nonrigid objects. The problem of registering a sequence of images of a nonrigid observed scene was addressed in [15]. The sequence of images were treated as samples from a multidimensional stochastic time series (e.g., an autoregressive model) which is learned. The stochastic model can then be used to extend the video sequence arbitrarily in time.

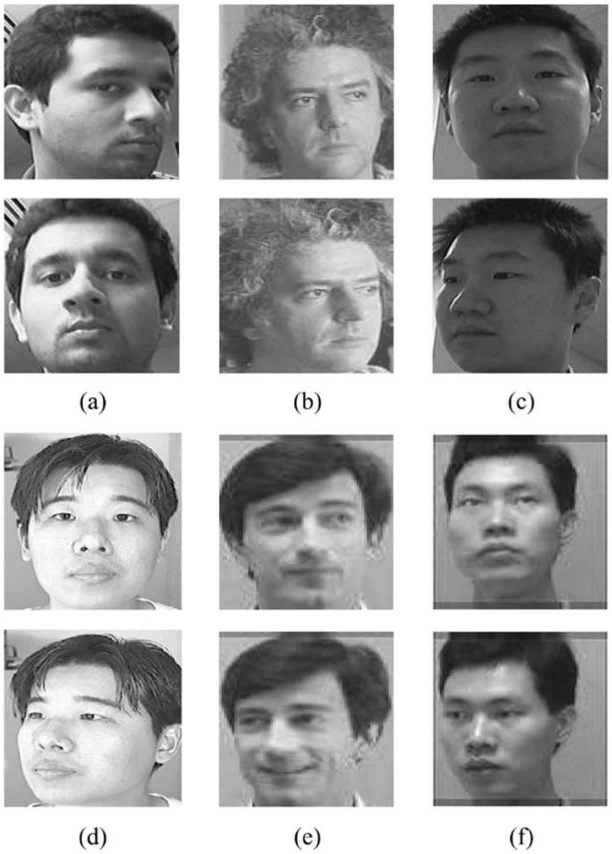


Fig. 1. Front and side views of the subjects 1–6 in our experiments. (a)–(c) Three subjects in the training set, while (d)–(f) represent three of the subjects of the test set.

B. Overview of Our Approach

The above methods for establishing correspondence rely, in essence, on matching of image tokens across groups of images. However, extraction of such image tokens (like the intensity or shape of significant features) is an inherently noisy process and most methods will be susceptible to error. In addition, it is extremely difficult to compute quantities which are invariant under different imaging conditions; both intensity and shape, the two most easily obtainable characteristics in an image, are dependent on the viewing angle. In this paper, we show that the availability of prior data in the form of a video sequence can help in developing robust correspondence schemes. In most application domains, obtaining this prior data is not a problem; e.g., for faces, it involves learning some general facial characteristics from a few video sequences of different faces.

The method presented here works with the edge image of local features (which gives an approximate notion of the 2-D shape of that feature), rather than their intensity. A doubly stochastic matrix, representing the probability of match between the features, is obtained using Sinkhorn normalization [16] and the prior information. A statistically optimal technique is proposed, which relies on minimizing the probability of error of a mismatch or equivalently maximizing the posterior density of the match given one of the features. The method works by matching the entire constellation of features in the two sets. The search space is no longer the set of features, but all their permutations ($N!$ for N features). The motivation for this *global* strategy (as opposed to the correspondence of individual features, that are local to that region) is that it emphasizes the "structural description of the pattern" [17] of the features. Use of prior information of the shape is an essential part of the scheme. The prior information is extracted from the video sequence in the form of an average representation of the features. The incorporation of prior information into the design of the detection strategy leads to a robust algorithm. The prior information can be collected once for different classes of objects and used across different objects in that class, e.g., in our application, the prior information can be collected once from video sequences of one or more faces and used across a large number of faces with similar characteristics. For 3-D face model generation, we learn the mean shape of a few significant features located on the face. The general shapes of eyes, nose, and lip features vary little from person to person, and thus a sufficient average shape may be obtained using data extracted from a few images of people over a range of viewing angles. Also, since the shapes of the different features are very different, considering their spatial arrangement in the face reduces any errors even further. A two-step optimization process is adopted, which consists of identifying occlusions followed by a probabilistic matching for each permutation of the two sets of features. It is also shown that, in practice, the search set can be made less than $N!$.

The above principles are used to obtain holistic 3-D models of a face from its video sequence by first creating partial models. The generation of 3-D face models is of particular importance to applications in multimedia, computer graphics and surveillance. In multimedia, 3-D face models can be used in video conferencing applications for efficient transmission. In computer graphics applications, 3-D face models form the basic building block on which facial movements and expressions can be added. Being able to build these models automatically from video data would greatly simplify such animation tasks where models are now built painstakingly with significant human intervention. In surveillance applications, 3-D models can be used for recognition across wide changes in viewing angles.

This paper is organized as follows. In Section II, we present our method to compute the probabilities for matching the individual features. Section III explains how to incorporate the spatial confguration of the features into the matching scheme. The correspondence algorithm is described in Section IV. The results of our algorithm applied to the problem of creating holistic 3-D models from partial ones is presented in Section V.

II. Registration Using Prior Models

A. Formulation of the Registration Problem

Our aim is to obtain correspondences between two sets of features extracted from images taken from different viewing directions and represented as sets of random variables, $\mathbf{X} = [X_1, \ldots, X_P]$ and $\mathbf{Y} = [Y_1, \ldots, Y_M]$. Each of the elements of the sets represents an image which is a collection of corners in a local region around the feature of interest, thus giving an idea of the 2-D shape of the region; hence, we use the term *shape cues*. Examples of these images can be seen in Fig. 5. Though the shapes of different features are usually significantly different, and therefore easier to match, they are often dependent on the viewing angle and their extraction process is extremely sensitive to noise. To overcome this, we use priors, which are the mean shape of each feature ("mean feature") collected from the video sequence over a range of viewing angles. Since the shapes of the features do not vary drastically for different people, the prior information can be collected only once and used across different video sequences.

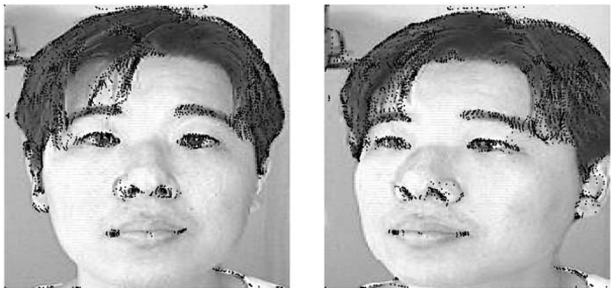


Fig. 2. Result of the corner finder algorithm on two images is represented by the small dots.

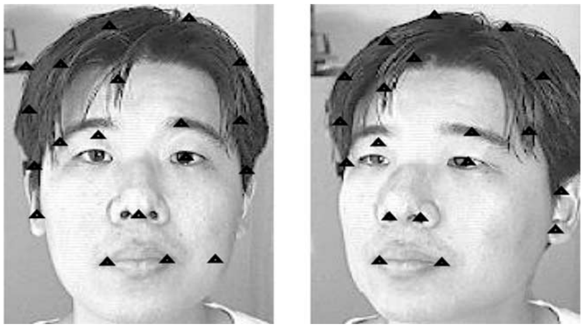


Fig. 3. Features identified in the front and side view images by applying a k-means clustering to the output of the corner-finder.

B. Computing the Feature Correspondence Probabilities

Let $\mu = \mu_1, \ldots, \mu_K$ represent the prior information of K features. Let H_i be the hypothesis that Y_i matches X; we wish to compute the *a posteriori* probability $P(H_i|X)$. Defining the event $\mathcal{E}_{X\mu_j} = \{X \text{ matches } \mu_j\}$, we hypothesize that the probability of X matching μ_j is directly proportional to the inner product of X with μ_j (since the inner product gives a measure of similarity). Since X and μ_j are images with nonnegative pixel values, the inner product will always be nonnegative. Then

$$P(\mathcal{E}_{X\mu_j}|X = X_n) = \frac{1}{\sum\limits_{i=1}^{K}\langle X_n, \mu_i\rangle}\langle X_n, \mu_j\rangle \tag{1}$$

where $\langle\cdot\rangle$ denotes inner product. For two images of size $P \times Q$, $\langle X_n, \mu_j\rangle = (1/PQ)\sum_{p=1}^{P}\sum_{q=1}^{Q} X_n(p,q)\mu_j(p,q)$. Similarly, the probability that Y_i matches X given the event $\mathcal{E}_{X\mu_j}$ is proportional to the inner product of Y_i and μ_j,

$$P(H_i|X, \mathcal{E}_{X\mu_j}) = \frac{1}{\sum\limits_{k=1}^{K}\langle Y_i, \mu_k\rangle}\langle Y_i, \mu_j\rangle. \tag{2}$$

Then, from the theorem of total probability [18], the *a posteriori* probability (which is the probability of X_n matching Y_i) is

$$P(H_i|X = X_n) = \sum_{k=1}^{K} P(H_i|X, \mathcal{E}_{X\mu_k})P(\mathcal{E}_{X\mu_j}|X = X_n). \tag{3}$$

The probabilities are represented in the form of a posterior probability matrix $\mathbf{P}(\mathbf{X}, \mathbf{Y})$. Our method works by maximizing the posterior probabilities. Viewed from a Bayesian perspective, this is equivalent to minimizing the Bayes risk, which is the

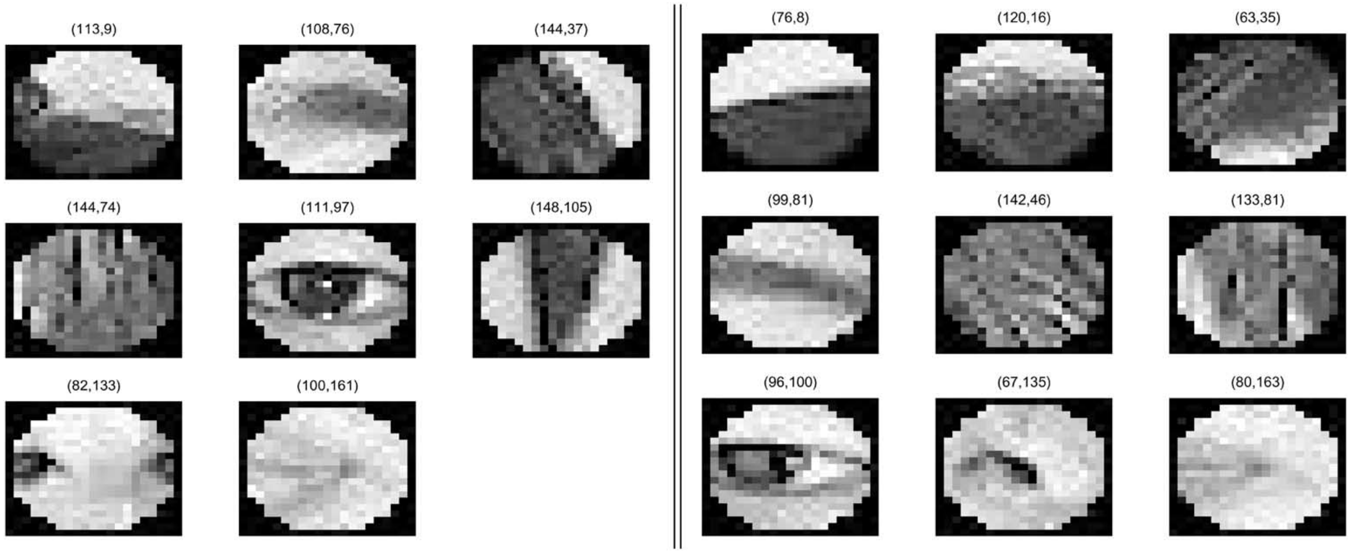


Fig. 4. Intensity blocks around the features to be matched in the front and side view. The numbers represent the position of the corresponding feature in the image.

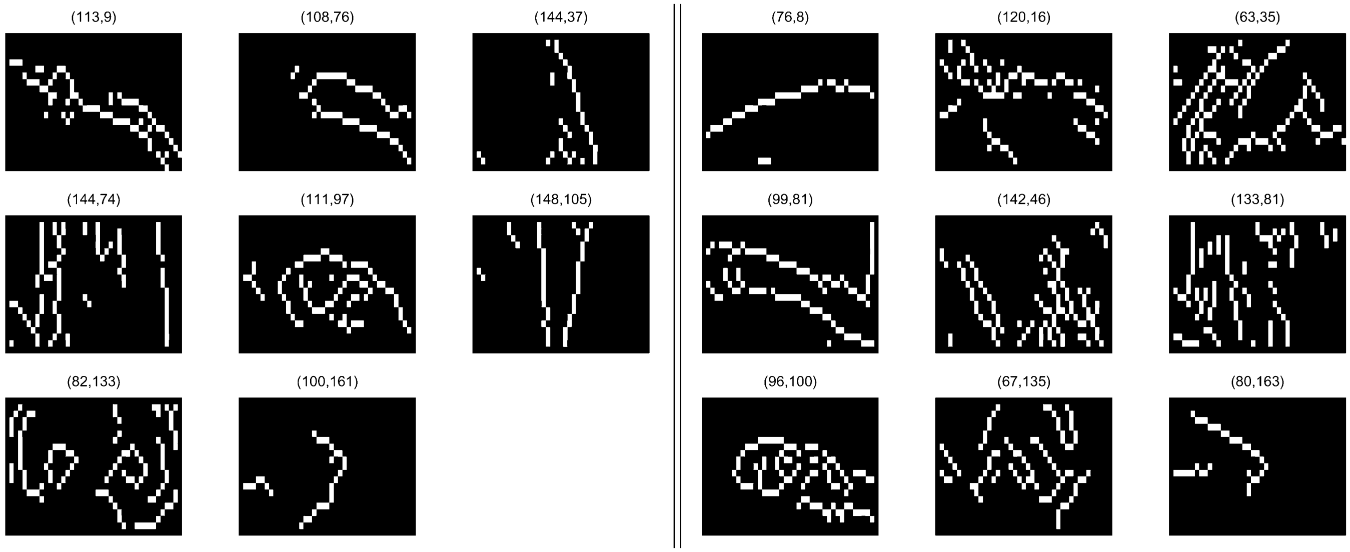


Fig. 5. Shape of the significant image attributes in the front and side views around the feature point whose position in the original image is indicated on top.

probability of error under the condition that incorrect decisions incur equal costs [19].

C. Prior Information

Assume that a feature $X_n(l)$[1] is corrupted by independent, zero-mean, additive noise ν. Let

$$X_n(l) = S_n(l) + \nu_n(l), \quad l = 1, \ldots, L \tag{4}$$

where $S_n(l)$ is the true unknown value of the feature. Then $\mu_n = E[X_n] = E[S_n] = (1/L(n)) \sum_{l=1}^{L(n)} X_n(l)$, since the noise is zero-mean and independent of the parameter, and the mean is computed over a range of viewing angles $L(n)$ ($L(n)$ can be different for different features). Thus we can compute the probability of a feature X_n in one image matching another feature Y_m in another image from (3). The probability is maximum when both X_n and Y_m match a particular prior feature μ_j.

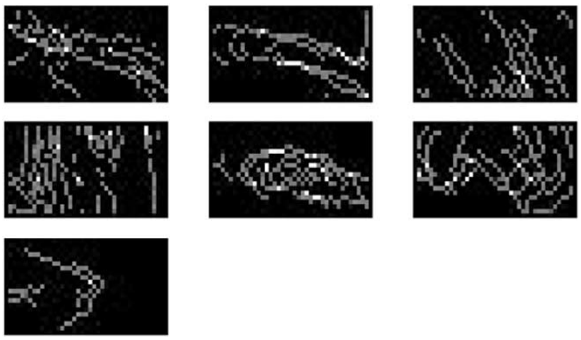


Fig. 6. Prior information (the shape representation averaged over a large number of viewing angles) which was precomputed.

[1] The notation $X_n(l)$ represents the image within a bounding box around the nth feature from the lth viewing position.

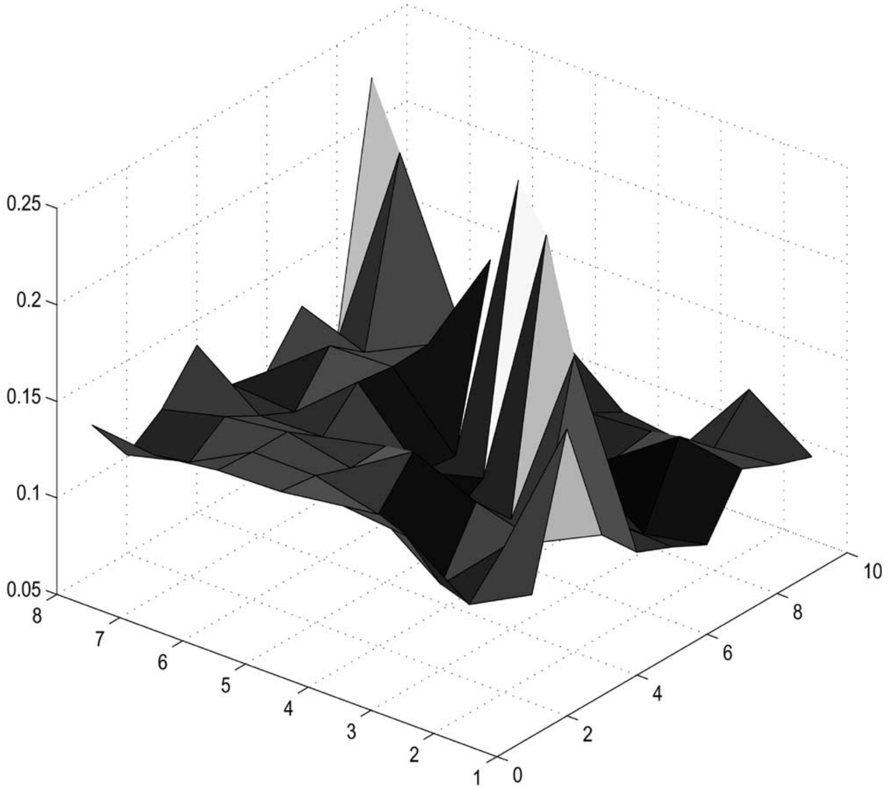


Fig. 7. Posterior density matrix.

D. Identifying Unpaired Features

In matching features from two different views, it is important to identify features present in one view but not in the other. If a particular feature X_n does not correspond to any feature in the set $\mathbf{Y}$, then $P(H_i|X = X_n)$, $i = 1, \ldots, M$ will not have any distinct peak (defined as the maximum whose difference with the second largest value exceeds a predefined threshold) and X_n can be identified. Similarly, if H'_i is the hypothesis that X_i matches Y, $P(H'_i|Y = Y_m), i = 1, \ldots, N$ will have a relatively flat profile if Y_m does not have a corresponding match in $\mathbf{X}$.

E. Correspondence Matrix

From the posterior probabilities, we would like to obtain a single doubly-stochastic matrix $\mathbf{C}(\mathbf{X}, \mathbf{Y})$, each row of which denotes the probability of matching the elements of $\mathbf{Y}$ given a particular $\mathbf{X}$, and each column the probability of matching the elements of $\mathbf{X}$ given a particular $\mathbf{Y}$. This is done by using the Sinkhorn normalization procedure to obtain a doubly-stochastic matrix by alternating row and column normalizations [16].

The advantage of using the Sinkhorn normalization procedure is that it allows us to use either $\mathbf{X}$ or $\mathbf{Y}$ as the reference feature set. It requires *a-priori* identification of unpaired features. This reduces the number of features that need to be matched and hence the combinatorics of the problem. As explained previously, the unmatched features are identified from their relatively flat probability profile. This is perfectly feasible since, as shown in the experiments in Fig. 9, the posterior probabilities always have a relatively flat trend for the case of unmatched features.

III. Matching the Spatial Arrangement of Features

Rather than computing a probability of match for individual features, a more reliable correspondence can be obtained if we consider the entire set of features, taking into account their relative spatial arrangement in the object, i.e., the constraints on the relative configuration of the features. Consider, for the purposes of this analysis, two sets of features $\mathbf{X}$ and $\mathbf{Y}$ having the same cardinality, say N (after identifying the unpaired features). We want to assign a probability of match of $\mathbf{X}$ against all possible permutations of $\mathbf{Y}$. Let the permutations of $\mathbf{Y}$ be represented by $\mathbf{Y}^1, \ldots, \mathbf{Y}^{N!}$, with $\mathbf{Y}^i = [Y_{(1)}, \ldots, Y_{(N)}]$, where $[Y_{(1)}, \ldots, Y_{(N)}]$ represents an ordering of $[Y_1, \ldots, Y_N]$. Let H^i represent the hypothesis that $\mathbf{Y}^i$ matches $\mathbf{X}$ (note the superscript used to distinguish the hypothesis for individual features). Then

$$P(H^i|\mathbf{X}) = \Pi_{j=1}^{N} P(H_{(j)}|X_j), \tag{5}$$

where $H_{(j)}$ is the hypothesis that $Y_{(j)}$ matches X_j for a particular permutation $\mathbf{Y}^i$. This assumes the conditional independence of each hypotheses H_j. This is a valid assumption for facial features when the change in expression is small; however, for other examples such as matching human body parts while in motion, this assumption would not hold since some body parts usually move together. Computing each of the probabilities in (5), we see that $P(H^i|X)$ is maximum when the permutation $\mathbf{Y}^i$ matches the set $\mathbf{X}$, element to element. In spite of considering all the permutations of one of the feature sets, the combinatorics of the problem is not high. This is because we are matching the

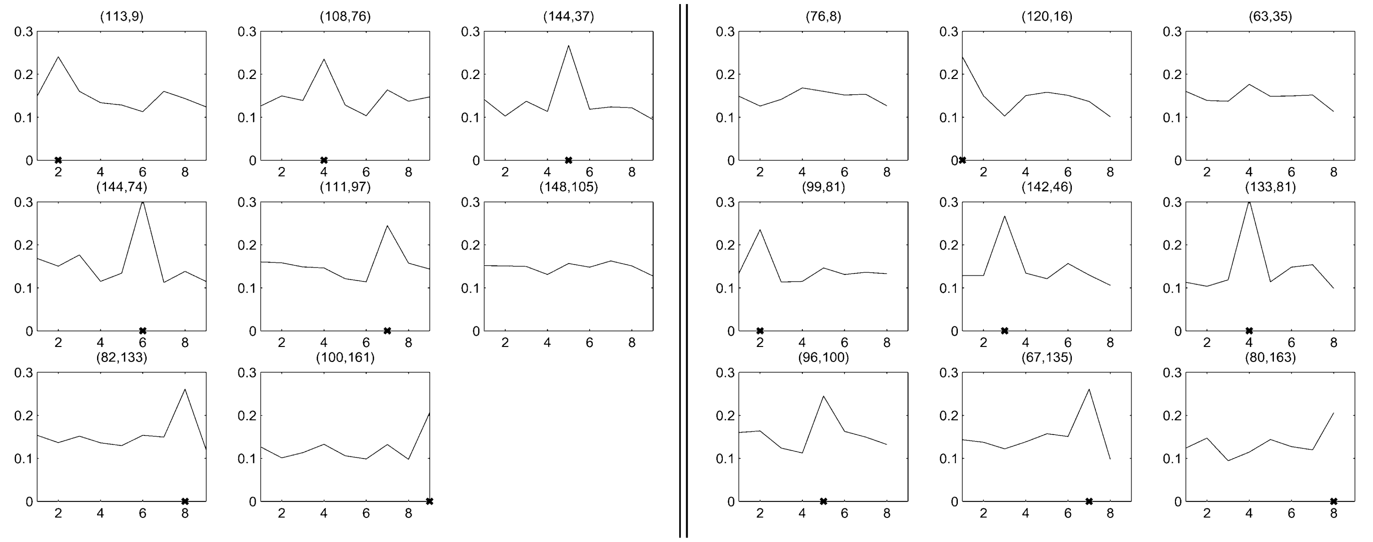


Fig. 8. *A posteriori* probabilities for each of the features in the front image and the side image, obtained respectively from each of the rows and columns of the correspondence matrix.

image in a region around a feature point of interest, and for the face there are usually only a few significant regions (e.g., eyes, nose, lips, etc.). In our experiments, we performed the matching using less than ten facial regions.

IV. The Correspondence Algorithm

We are given two images $\mathcal{I}_1$ and $\mathcal{I}_2$, and the precomputed prior information $\mu_1, \ldots, \mu_K$.

1) *Feature Extraction*: Compute the set of features $\mathbf{X} = [X_1, \ldots, X_P]$ and $\mathbf{Y} = [Y_1, \ldots, Y_M]$ using a suitable feature extraction method (in our case, a corner-finder algorithm).
2) *Compute Probability of Match*: Compute the match probabilities from (3) using the prior information $\mu_1, \ldots, \mu_K$.
3) *Identify Unpaired Features*: Identify those features present in one view, but not in the other as explained above. At the end of this process, we are left with two sets with the same cardinality (denoting the paired features) which have to be matched. Denote them by $\mathbf{X} = [X_1, \ldots, X_N]$ and $\mathbf{Y} = [Y_1, \ldots, Y_N]$.
4) *Sinkhorn Normalization*: Compute the correspondence matrix $\mathbf{C}(\mathbf{X}, \mathbf{Y})$ by applying the Sinkhorn normalization procedure to the match probabilities after removing the unpaired features.
5) *Compute the Probability of the Spatial Arrangement of the Features*: Compute the posterior probability for matching $\mathbf{X}$ with all permutations of $\mathbf{Y}$, i.e., $P(H^i|\mathbf{X})$, $i = 1, \ldots, N!$ from (5).
6) *Search for Best Match*: Obtain $i = \arg\max_i P(H^i|\mathbf{X})$. Assign $\mathbf{Y}^i = [Y_{(1)}, \ldots, Y_{(N)}]$ as the match to $\mathbf{X}$.

Reducing the Search Space: The search space in the last step of the above algorithm is of size $N!$. In practice, the search space can be reduced. For each $X = X_n, n = 1, \ldots, N$ for the paired sets of features, identify the set $\bar{Y}_n = \{Y_i : P(H_i|X = X_n) > p\}$, where p is an appropriately chosen threshold. Alternatively, we can choose the $\{Y_i\}$ that have the largest l values of the posterior densities. This smaller set identifies those features in $\mathbf{Y}$ which are the closest to a particular feature in $\mathbf{X}$. We can then compute the probability of match for the permutations of $\mathbf{Y}$ in this reduced set. The actual number of elements contained in the search space will depend on the exact values of the probabilities of $\bar{Y}_n, n = 1, \ldots, N$.

V. Experimental Analysis and Applications

We present the results of our algorithm applied to the problem of registering two images of a face taken from two different viewing directions. We use a database consisting of 24 people whose images have been obtained under different imaging conditions and who bear widely varying facial features. We present the results of the probabilistic correspondence algorithm for each of these subjects and the result of the global alignment strategy for a few of them. Finally, we show how our registration algorithm can be used for building holistic 3-D models from partial ones.

The database of test subjects is explained in Table I. The images of the first six subjects are shown in Fig. 1, with both front and side views. The images of the other subjects are not shown in the paper on their personal request. The images were obtained from a database available on the World Wide Web at http://images.ee.umist.ac.uk/danny/database.html. Details can be found in [20] and the data can be viewed by the interested reader at the website.

The prior information was precomputed from the video sequences of the first three subjects (1, 2, and 3) in Table I [Fig. 1(a)–(c)]. We will refer to these subjects as the training set. The remaining subjects will be referred to as the test set. Before we proceed to present the results on this entire dataset, we will present the details of our algorithm on Subject 4 (the first in the test set). The details will be similar for the other subjects, and hence we present only the final results.

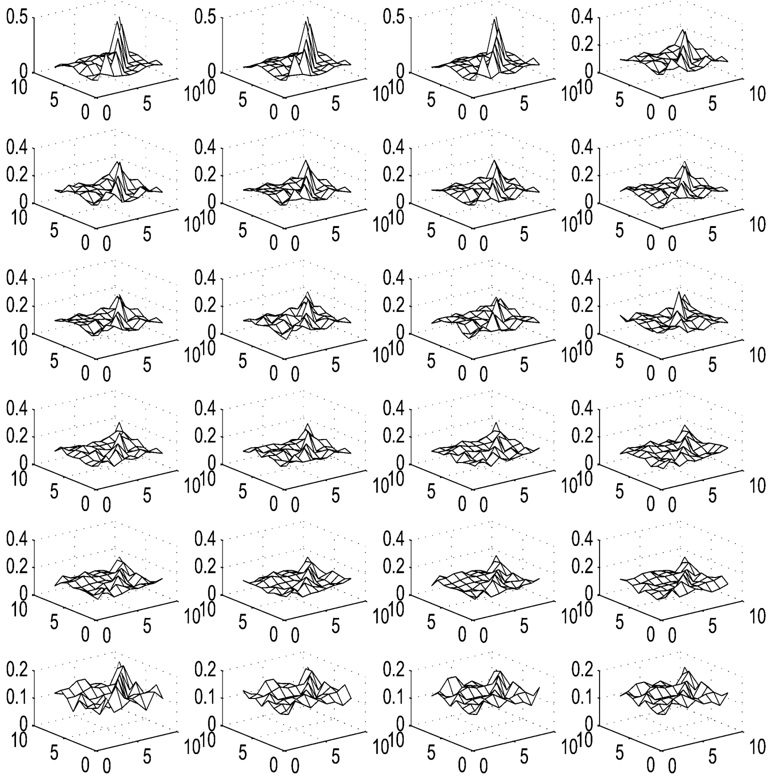


Fig. 9. Probability matrices obtained for subjects 1–24 in our experiment. The subjects are arranged in row-major order, i.e., the numbering of the subjects increases sequentially across each row.

A. Feature Selection and Prior Extraction

To select the features that need to be registered, we use a corner finder algorithm based on an interest operator[2] [21]. Fig. 2 shows the output of the corner finder algorithm represented by the small dots. Given this output defining the corners of the image, a clustering algorithm, like k-means, was used to identify feature points. The k-means algorithm computes the centroids of these dots and identifies them as the important features on the face. The local images, formed by the dots around the features, need to be matched. The k-means algorithm is thus used to filter out spurious points in the output of the corner finder algorithm; a few important clusters are identified and then only the points around these clusters are retained. It is very important to understand that we match entire local regions around these feature points, not just the points. Hence only a few such regions (less than ten) are enough, since there are only a few distinct aspects of a face. Fig. 3 plots two sets of features identified using this strategy. However, in order to avoid the feature matching problems that can arise due to the symmetry of a face, we only considered features located in the right 70% of the original images. In addition, features lying in the region near the image boundaries were neglected. We will present our results on this smaller set of features. Fig. 4 plots the intensities in the local regions around the features and Fig. 5 plots the output of the corner-finder algorithm representing the 2-D shape around these features. Fig. 6 represents the precomputed prior information in the form of the mean features. The prior was collected by tracking a set of features across multiple frames of the video sequences of subjects 1, 2, and 3 and then integrating them out. These subjects were chosen because they had significantly different facial characteristics and thus covered a large class of features.

B. Estimation of Posterior Probabilities

Fig. 7 gives a graphical representation of the posterior probability matrix $\mathbf{P}(\mathbf{X}, \mathbf{Y})$ obtained before the Sinkhorn normalization procedure. It can be seen that there is a distinct peak for each row and column of the matrix, corresponding to matching of a pair of features. A distinct peak is defined as the maximum of the probability values in that row or column and whose difference with the second largest value is above a certain threshold. The valleys of this surface plot, representing rows or columns with no peaks, correspond to unmatched pairs of features. Fig. 8 plots the rows and columns of $\mathbf{P}(\mathbf{X}, \mathbf{Y})$ respectively. The true values (as obtained manually) are marked by a $*$ on the horizontal axis, except for those which are unmatched (the unpaired features).

C. Matching the Spatial Arrangement of Features

Fig. 11(a) plots the probabilities for matching $\mathbf{X}$ against all possible permutations of $\mathbf{Y}$. Comparison with Fig. 8 shows that there is a very distinct peak in this case, justifying our earlier assertion that taking into account the spatial arrangement of the features leads to a more robust algorithm. Since there are only a few regions to match in the two views, the combinatorics of the problem of matching all arrangements is not a problem.

[2] The interest operator computes the matrix of second moments of the local gradient and determines corners in the image based on the eigenvalues of this matrix.

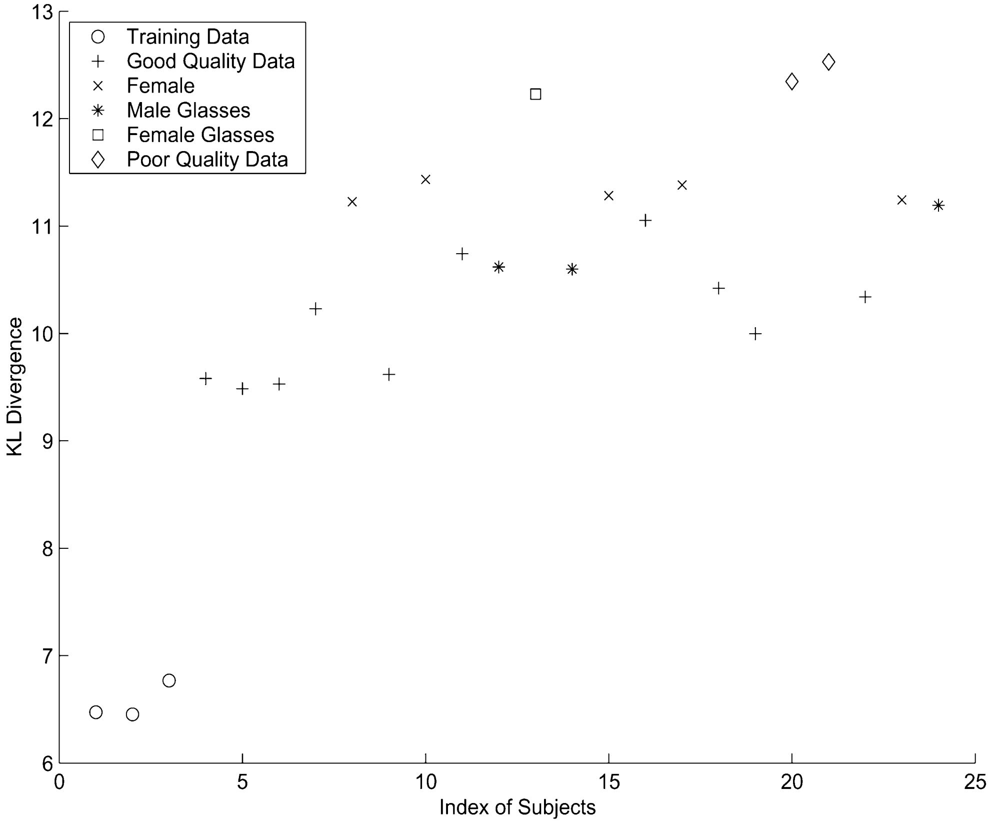


Fig. 10. KL divergence between the obtained probability distribution and the ideal one for all the subjects in the experiments.

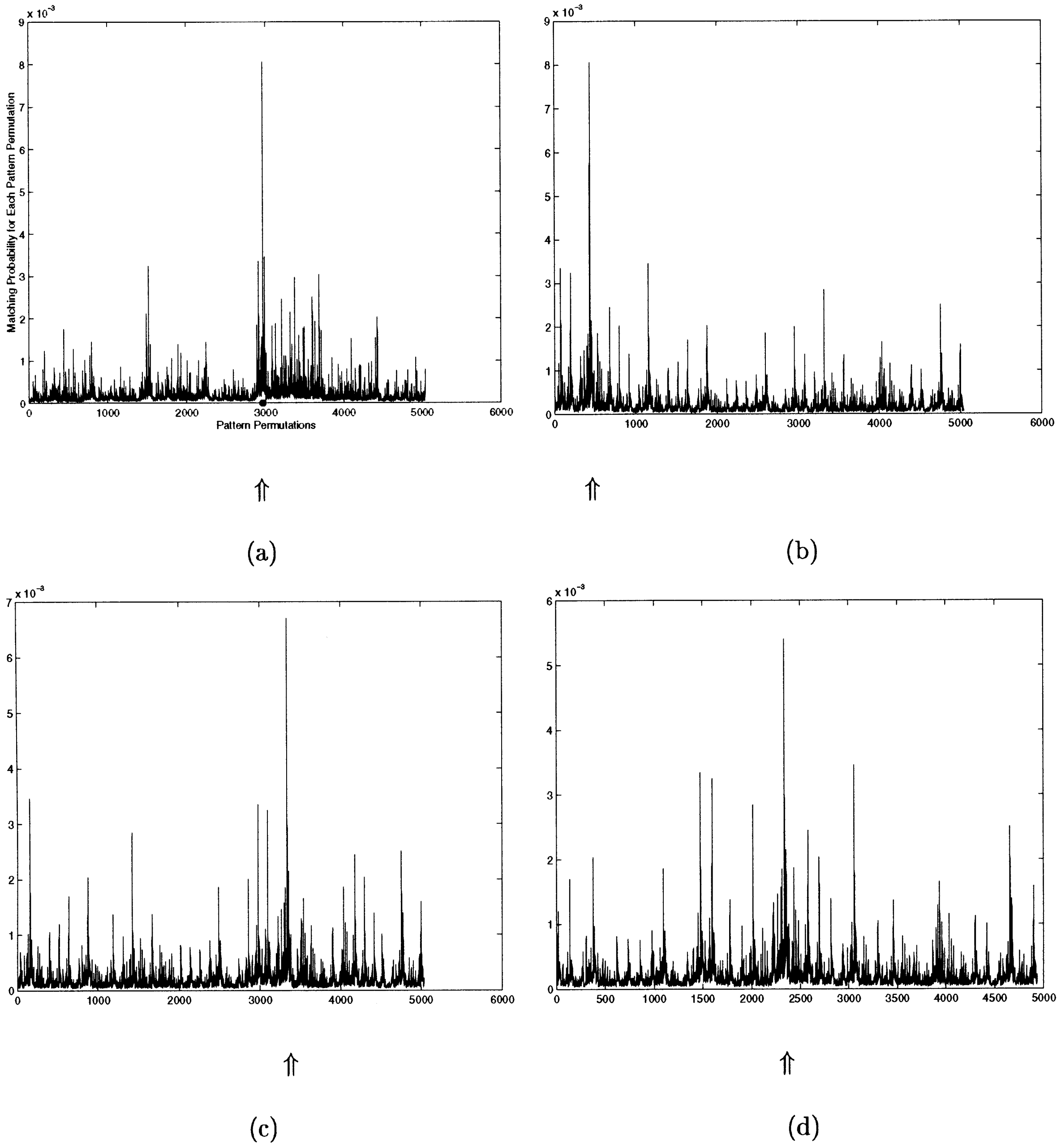
D. Results on Complete Dataset

Having explained the details of our algorithm on one particular example, we present the result of applying our method to the dataset of 24 individuals described above. Analysis of the images in the dataset suggests that they have widely different characteristics, e.g., different ethnic backgrounds, different gender, with or without eyeglasses, with or without beard or moustache, different imaging conditions etc. Thus it is to be expected that the results of the probabilistic matching technique would be different. The probability matrices for each of the subjects is shown in Fig. 9. For easy comparison between the different matrices, we tried to keep the numbering of the features the same (e.g., eye in always number 5 in the front view). This is done manually and is not an essential part of the algorithm. However, in some cases it was not possible because of the kind of features identified. Comparison of these plots with Table I shows that the distinctness of the peaks in the probability matrices does indeed decrease as the features of the test set move farther away from those of the training set. In order to get a quantitative feel of the deviation of the probability matrix from the ideal one, we compute the Kulback–Leibler (KL) divergence [22]. The ideal matrix is the one that would be obtained if the match was perfect. It contains a 1 for the correct match in each row and zeros elsewhere and is defined manually. The values of the KL divergence are tabulated in Table I and plotted in Fig. 10. Analysis of the KL divergence reveals how the performance of our algorithm degrades as a function of the facial characteristics.

Surely, by themselves, the probability matrices are not enough to identify all the corresponding features. It is in these cases that the global matching scheme using the spatial arrangement of all the features is most important. In Fig. 11(b)–(d), we present the result of the spatial arrangement for the three subjects having the highest KL divergence values, namely 13, 20, and 21. It can be seen that there is a distinct peak in the probabilities in all three cases, thus proving that our method is indeed robust and can be applied to a large number of examples.

E. Importance of Prior Information

We now demonstrate the importance of the prior information, again resorting to our special example of subject 4. In Fig. 12, we plot the probabilities of match of each feature in $\mathbf{X}$ against the different features in $\mathbf{Y}$, where we do not have the precomputed prior information. The probabilities were estimated using the shape similarity between the two features. This was done using the standard technique of computing the ratios of the eigenvalues of the first and second central moments of the coordinates of the set of points representing the features [21]. This was extended to consider the permutations of the features so as to take advantage of the global arrangement. Fig. 13 plots the probability of matching the spatial arrangement of the features without the advantage of the prior information. In both these cases, we see that the peaks of the probabilities do not correspond to the true match, as indicated in the plots. This emphasizes the importance of the prior information and shows how a simple correlation based matching technique can be modified to provide a very robust solution by incorporating suitable information gathered from the video data.



(a)

(b)

(c)

(d)

Fig. 11. Probability of matching $\mathbf{X}$ against all permutations of $\mathbf{Y}$. The true value is marked with a ⥣ below the horizontal axis. The plots are for subjects 4, 13, 20, and 21, arranged in a row major order.

F. Application to 3-D Model Alignment

We now demonstrate the application of our correspondence algorithm for aligning two partial models of a human face obtained from different views. The models were obtained from a video sequence of a person moving his head in front of a static camera using structure from motion (SfM) [23], [24]. The video sequence was split into two portions, corresponding to the front and side views of the face. The two partial models were obtained from these two portions of the video sequence. In order to obtain the 3-D models from video, a set of features were tracked and the depth and camera motion at these points were computed using a multiframe structure from motion (SfM) algorithm [25]. The SfM algorithm worked by fusing the depth estimates obtained from two images using optical flow techniques. The fusion was done using robust statistics and a generic model of a face. The error in the reconstruction was estimated and compensated for. Details of the 3-D modeling algorithm are available in [26]. Fig. 14 depicts the two models, one from the front, the other from the side, which we aim to integrate into one holistic model.

In order to align these two partial models, one image, obtained from each of the views, is considered and our algorithm is used to obtain correspondence between the features automatically selected in these images. Prior information for important

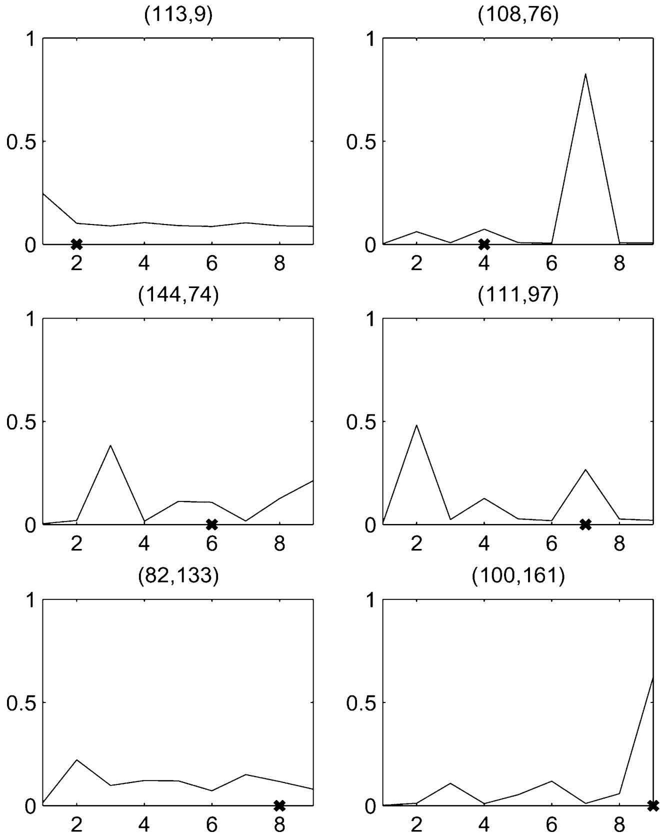


Fig. 12. Probability of match for each of the features in the front image, for the case where prior information is not available.

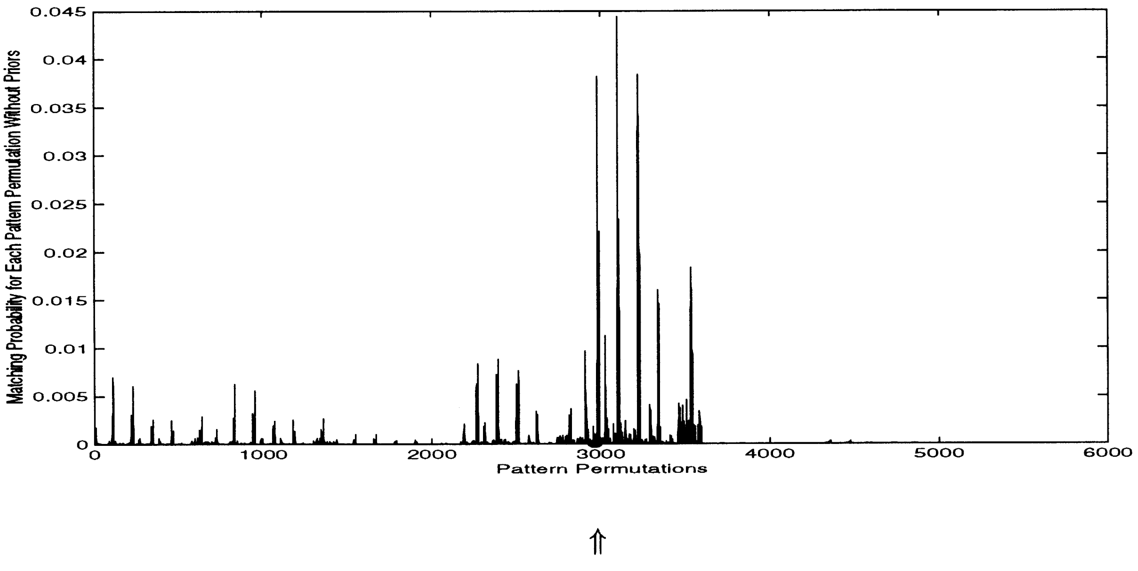


Fig. 13. Probability of match for the shape of each feature in the front image against all possible combinations of the features in the side view, for the case where prior information is not available. The true value is marked with a ⇑ below the horizontal axis.

features in a human face was precomputed, and used for this application (as explained earlier). Our algorithm presented in Section IV was then used to obtain the correspondences between tefhe different features. Having obtained the feature correspondence, we compute the local affine transformation between the two models for each of the features separately, i.e., $\mathbf{y}_i = \mathbf{R}_i\mathbf{x}_i + \mathbf{T}_i$ where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the 3-D coordinates of a matching pair of points and $\mathbf{R}_i$ and $\mathbf{T}_i$ the rotation and translation for a local region around the feature i. Fig. 14 also shows two views of the complete model after alignment. Our feature

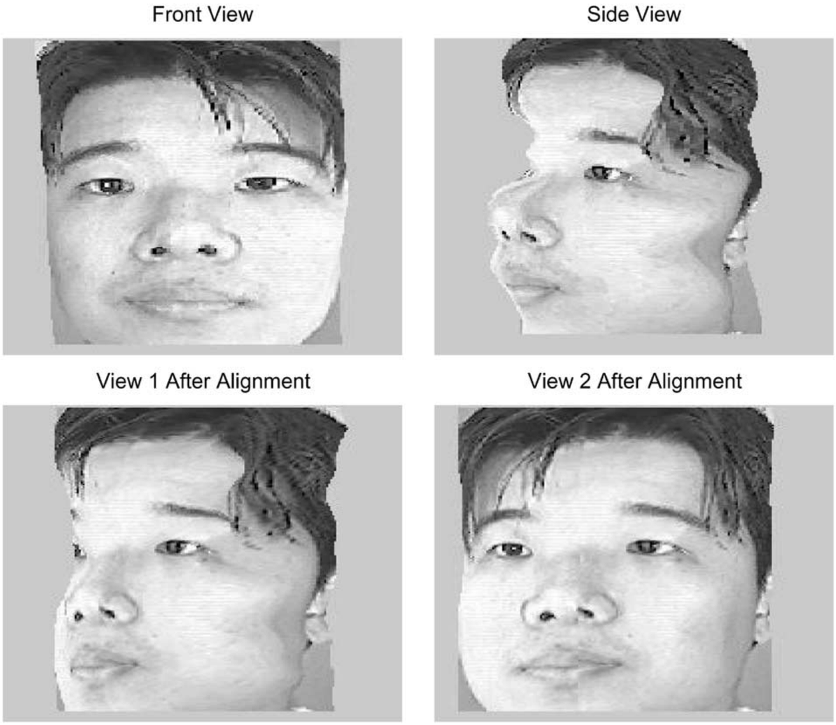


Fig. 14. The 3-D partial models from the front and side which are used as input to the algorithm are shown in the top row and two views of the 3-D model obtained after the alignment are shown in the bottom row.

correspondence algorithm can also be used to obtain good initial conditions for precise registration methods described in [1] and [13].

VI. Conclusion

In this paper, we have presented a probabilistic framework for matching two sets of features, extracted automatically from images, which takes into consideration the global structure of the feature sets. The Sinkhorn normalization procedure is used to obtain a doubly stochastic matrix denoting the probabilities of match for the two feature sets. The method works by minimizing the probability of a mismatch (using the Bayes error criterion) between the shapes of the features, after taking into account their spatial arrangement. Robustness is achieved by including prior information regarding these feature sets. We emphasize that the prior can be easily obtained from video, and needs to be computed only once for a class of objects. An application of this method to 3-D model alignment of a human face was demonstrated.

Acknowledgment

The authors would like to thank Dr. S. Teller of MIT and Dr. A. Papamarcou and Dr. A. Rosenfeld at the University of Maryland for their comments and suggestions.

References

[1] P. Besl and N. McKay, “A method for registration of 3D shapes,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 239–256, Feb. 1992.

[2] A. W. Fitzgibbon. Robust registration of 2D and 3D point sets. presented at British Machine Vision Conf.. [Online]. Available: http://www.robots.ox.ac.uk/ vgg

[3] B. Vemuri and J. Aggarwal, “3D model construction from multiple views using range and intensity data,” in *Proc. Computer Vision and Pattern Recognition Conf.*, 1986, pp. 435–437.

[4] G. Blais and M. Levine, “Registering multiview range data to create 3D computer objects,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 820–824, Aug. 1995.

[5] P. Beardsley, P. Torr, and A. Zisserman, “3D model acquisition from extended image sequences,” in *Proc. Eur. Conf. Computer Vision*, 1996, pp. 683–695.

[6] R. Koch, M. Pollefeys, and L. Van Gool, “Multi viewpoint stereo from uncalibrated sequences,” in *Proc. Eur. Conf. Computer Vision*, 1998, pp. 55–71.

[7] J. VandenWyngaerd, L. VanGool, R. Koch, and M. Proesmans, “Invariant-based registration of surface patches,” in *Proc. Int. Conf. Computer Vision*, 1999, pp. 301–306.

[8] C. Schmid, R. Mohr, and C. Bauckhage, “Comparing and evaluating interest points,” in *Proc. Int. Conf. Computer Vision*, 1998, pp. 230–235.

[9] C. Tomasi and J. Shi, “Good features to track,” *IEEE Comput. Vis. Pattern Recognit.*, pp. 593–600, 1994.

[10] T. Cham and R. Cipolla, “A statistical framework for long-range feature matching in uncalibrated image mosaicing,” *IEEE Comput. Vis. Pattern Recognit.*, pp. 442–447, 1998.

[11] F. Badra, A. Qumsieh, and G. Dudek, “Robust mosaicing using Zernike moments,” *PRAI*, vol. 13, no. 5, p. 685, Aug. 1999.

[12] N. Ritter, R. Owens, J. Cooper, R. Eikelboom, and P. Van Saarloos, “Registration of stereo and temporal images of the retina,” *IEEE Trans. Med. Imag.*, vol. 18, pp. 404–418, May 1999.

[13] P. Viola and W. Wells, III, “Alignment by maximization of mutual information,” *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, September 1997.

[14] M. Burl, M. Weber, and P. Perona, “A probabilistic approach to object recognition using local photometry and global geometry,” in *Eur. Conf. Computer Vision*, 1998.

[15] A. W. Fitzgibbon. Stochastic rigidity: Image registration for nowhere-static scenes. presented at Int. Conf. Computer Vision. [Online]. Available: http://www.robots.ox.ac.uk/ vgg

[16] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *Ann. Math. Statist.*, vol. 35, pp. 876–879, 1964.

[17] K. Fu, *Syntactic Pattern Recognition and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[18] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1991.

[19] M. Srinath, P. Rajasekaran, and R. Viswanathan, *Introduction to Statistical Signal Processing With Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[20] D. Graham and N. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," *Face Recognition: From Theory to Applications*, ser. NATO ASI Series F, Computer and Systems Sciences, vol. 163, pp. 446–456, 1998.

[21] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[22] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[23] Z. Zhang and O. Faugeras, *3D Dynamic Scene Analysis*. Berlin, Germany: Springer-Verlag, 1992.

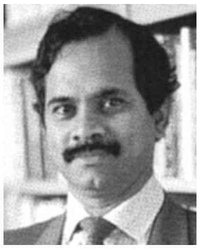
[24] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[25] R. C. A, "Statistical Analysis of 3D Modeling From Monocular Video Streams," Ph.D. dissertation, Univ. Maryland, College Park, 2002.

[26] A. R. Chowdhury, S. Krishnamurthy, T. Vo, and R. Chellappa, "3D face reconstruction from video using a generic model," in *Int. Conf. Multimedia and Expo.*, Lausanne, Switzerland, 2002.

Amit K. Roy-Chowdhury received the Ph.D. degree in 2002 from the Department of Electrical and Computer Engineering at the University of Maryland, where he worked on statistical error characterization of 3-D modeling from monocular video sequences.

He was then a Research Scientist at the Center for Automation Research, University of Maryland, College Park, MD, the Lead Scientist on projects related to human recognition and activity inference, with applications in surveillance, multimedia and communications. Since 2003, he has been with the Department of Electrical Engineering, University of California, Riverside. His research interests are in image and video processing, computer vision and statistical signal processing.



Rama Chellappa (S'79–M'81–SM'83–F'92) received the B.E. (Hons.) degree from the University of Madras, India, in 1975 and the M.E. (Distinction) degree from the Indian Institute of Science, Bangalore, in 1977, and the the M.S.E.E. and Ph.D. Degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1978 and 1981, respectively.

Since 1991, he has been a Professor of electrical engineering and an affiliate Professor of computer science at the University of Maryland, College Park. He is also affiliated with the Center for Automation Research (Director) and the Institute for Advanced Computer Studies (Permanent member). Prior to joining the University of Maryland, he was an Assistant (1981–1986) and Associate Professor (1986–1991) and Director of the Signal and Image Processing Institute (1988–1990) with the University of Southern California (USC), Los Angeles. Over the last 22 years, he has published numerous book chapters, peer-reviewed journal and conference papers. He has edited a collection of *Papers on Digital Image Processing* (Los Alamitos, CA: IEEE Computer Society Press), co-authored a research monograph, with Y. T. Zhou, on *Artificial Neural Networks for Computer Vision* (Berlin, Germany: Springer-Verlag, and co-edited a book on *Markov Random Fields*, with A.K. Jain (New York: Academic). He was co-Editor-in-Chief of *Graphical Models and Image Processing*. His current research interests are face and gait analysis, 3-D modeling from video, automatic target recognition from stationary and moving platforms, surveillance and monitoring, hyper-spectral processing, image understanding, and commercial applications of image processing and understanding.

Dr. Chellappa is now serving as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), and an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON NEURAL NETWORKS. He served as a member of the IEEE Signal Processing Society Board of Governors during 1996–1999. Currently, he is serving as the Vice President of Awards and Membership for the IEEE Signal Processing Society. He has received several awards, including NSF Presidential Young Investigator Award, an IBM Faculty Development Award, the 1990 Excellence in Teaching Award from the School of Engineering at USC, and the 1992 Best Industry Related Paper Award from the International Association of Pattern Recognition (with Q. Zheng), the 2000 Technical Achievement Award from the IEEE Signal Processing Society. He was elected as a Distinguished Faculty Research Fellow (1996–1998, 2003) at the University of Maryland. He is a Fellow of the International Association for Pattern Recognition. He has served as a General Technical Program Chair for several IEEE international and national conferences and workshops.

Trish Keaton is currently pursuing the Ph.D. degree in electrical engineering at the California Institute of Technology, Pasadena. She is also a Research Scientist at HRL Laboratories (formerly Hughes Research Laboratories), Malibu, CA, where she is the principal investigator of projects focused on 3-D human tracking and activity recognition with applications involving surveillance, ubiquitous and wearable computing. Her research interests include computer vision, robust tracking systems, human computer interfacing, graphical modeling, level set methods, and multimedia indexing and retrieval.