AUTOMATIC FACE RECOGNITION USING STEREO IMAGES

Anjali Samani, Joab Winkler, Mahesan Niranjan

The University of Sheffield, Department of Computer Science, Sheffield S1 4DP, UK
a.samani/j.winkler/m.niranjan@dcs.shef.ac.uk

ABSTRACT

Face recognition is an important pattern recognition problem in the study of natural and artificial learning systems. In typical optical image based face recognition systems, the systematic variability that arises from representing the three dimensional (3D) shape of a face by a two dimensional (2D) illumination intensity matrix is treated as a random variable, and it is obtained by collecting examples of faces in different poses with respect to the camera. More sophisticated 3D recognition systems employ specialist equipment (*e.g.* laser scanners) to measure the shape of the face, and they perform either pattern matching in three dimensions or they use projections from 3D models to match against 2D images. It is shown here that optical images obtained with a pair of stereo cameras may be used to extract depth information in the form of *disparity values*, and thereby significantly enhance the performance of a face recognition system.

1. INTRODUCTION

Pattern recognition involves learning a statistical model, either in the form of a parametric density (*e.g.* mixture of multivariate Gaussians) or in the form of a functional mapping (*e.g.* artificial neural networks) from a set of examples of different patterns, and subsequently making inferences on unseen examples. Variabilities inherent in patterns that belong to a particular class make this an interesting and challenging problem, both in an artificial intelligence setting and in the human perceptual learning setting. Variabilities arise for several reasons, including the inadequacy of the features chosen to represent patterns, contextual effects such as co-articulation in acoustic speech patterns, and random fluctuations due to sensor noise. The decomposition of these variabilities and the appropriate representation of uncertainties should be an essential part of the design of pattern recognition systems.

Photographs of faces are 2D images that capture pixel intensities resulting from an affine projection of an underlying 3D object on the camera image plane. Different poses of the face with respect to the camera result in a variation in the image space, which is systematic. This systematic variability is usually modelled as a random variable by collecting a number of images per subject, captured at different tilts of the face with respect to the camera axis.

3D face recognition is usually approached in one of three ways. Model-based approaches (*e.g.* [1]) use a generic 3D face model, created using the laser-scanned face models of all the subjects in the database. Although very accurate, this approach requires extensive subject co-operation and may require manual identification of fiducial points on the facial surface. Computational intensity and heavy reliance on pre- and post-processing limits the applications of such a system.

The second approach reconstructs the facial surfaces for each person in the dataset using precise depth values, which may be obtained from a variety of cues such as stereo images, range images, etc. Surface matching techniques (*e.g.* ICP) are used for identification. Stereo-based techniques of [2, 3, 4, 5] use this approach and report recognition accuracies of over 90%. However, they require accurately calibrated cameras to reconstruct the surfaces, making it difficult to deploy such a system out of laboratory conditions, where the cameras may be subject to perturbations and re-calibration may not be possible. Incorrect camera matrices lead to incorrect reconstructions, and hence identification.

In the $2\frac{1}{2}$D approach, the depth information is encoded directly in a 2D image by replacing the intensity values with the depth values, so that the new pixel values correspond to the surface geometry of the 3D object (e.g. range images, depth maps) [2, 6]. Such a representation captures the depth information of a scene, whilst still enabling all the existing 2D image processing and face recognition techniques to be used. A comprehensive survey of 3D and multi-modal systems, combining 3D shape and 2D texture, can be found in [7, 8].

We used a pair of digital cameras in a stereo setup to capture the depth information, and our experiments show that automatic face recognition can be significantly improved by using a combination of texture and depth information. In existing systems, recognition is preceded by camera-calibration, triangulation and surface reconstruction, all of which are error prone, but these operations are not performed in the work described in this paper. Wavelet transforms were used to solve the correspondence problem between the left and right camera images. This enabled a disparity field that contains information about depth, proportional up to parameters of the camera matrices (which are constant across all faces and hence do not play any role in discrimination), to be constructed. The dis-

parity matrix (Fig. 1) shows the displacement of each pixel between the two images. The resulting classifier has a single channel intensity image and a disparity matrix, and when they are normalised and appended to each other, the combination forms a high dimensional vector representation, a *composite image*, of the face that captures intensity and depth information. We design a statistical pattern recognition system in this space and test it with the state of the art Eigenface approaches. It is seen from Table 1 that a significant improvement in performance in the classification accuracy is observed.

2. DATA

We collected 540 pairs of stereo images of 22 individuals (Sheffield Dataset) using a pair of Olympus Camedia C-200Z cameras. Images with varying pose, illumination and expressions were captured (see Figure 2). The effects of "harsh lighting" were simulated in two of the images by illuminating the face strongly from the left side and from underneath. Individuals who wore glasses were photographed with and without glasses. The database also contains two females with a head-scarf, one of whom is also photographed with and without the scarf. Images of an individual with and without glasses/head-scarf are treated as belonging to different classes.

The subjects were seated about 60cm from a smooth monotonic background, approximately 300cm from the cameras. The cameras were separated by a horizontal baseline of approximately 22cm. To maintain uniformity in illumination across all individuals, fluorescent lights were used instead of natural lighting. Although this setup minimises shadows and reflections, special effort to control these effects was not made. The pose images were captured by asking the subject to face signs placed strategically around the room, such that the degree of rotation was not strictly controlled. In order to keep the dataset as realistic as possible, restrictions were not imposed on the expressions displayed by the subjects.

3. METHODOLOGY

The $2\frac{1}{2}$D images were computed using Magarey and Dick's complex wavelets based multiresolution stereo image matching algorithm [9]. It employs a coarse-to-fine matching strategy, and the disparity field estimated at each level of decomposition is refined and regularised by using the estimated disparity from the previous coarser level. The field is interpolated, scaled and propagated to the next finer level in order to obtain robust disparity estimates. The algorithm is summarised in pseudocode form in Figure 3, details on complex wavelets and mathematical exposition of the algorithm are in [10] and its application to matching face images is in [9].

The images in the 2D, $2\frac{1}{2}$D and the composite spaces are classified using Turk and Pentland's principal component analysis (PCA) based technique of Eigenfaces [11].

PCA is used in the Eigenfaces algorithm to find the vectors that best represent the distribution of face images in the face space. Each $(N \times N)$ image $\mathbf{\Gamma}_i$ is considered as a vector in N^2-dimensional space. The covariance matrix C is computed using the mean-subtracted images ($\mathbf{\Phi}_i = \mathbf{\Gamma}_i - \mathbf{\Psi}$), where $\mathbf{\Psi}$ is the mean face. A test image $\mathbf{\Gamma}_j$ is recognised by first transforming it into its Eigenface components. The weights ω_k form a vector $\Omega^T = [\omega_1, \omega_2, \ldots, \omega_{M'}]$ that describes the contribution of each Eigenface in representing the input image face, treating the Eigenfaces as a basis set for face images. See equations 1 and 7 in [11] for further details. The face classes Ω_l are calculated by averaging the results of the Eigenface representation over a small number of face images of each individual.

The class k of the test image is the one that minimises the distance between the vectors Ω and Ω_k. We used Craw's formulation of the Mahalanobis distance instead of the Euclidean, since it has better discriminatory power and is known to give superior results [12].

4. RESULTS

Tests were carried out on 256×256 2D greyscale images. The $2\frac{1}{2}$D images measured 128×128 and were generated using 7 levels of complex wavelet decomposition. The intensity values and the disparity values were normalised so that they lay between 0 and 1.

Subsets of 30, 150, 300 and 539 (leave-one-out cross validation) images were used to train the Eigenfaces classifier. This yielded 1, 5, 10 and approximately 17 images per class for training, and the rest were used to test the classifier. In order to obtain reliable measures of classifier accuracies and error measures, the recognition experiments were run 10 times (except the leave-one-out cross validation). A different set of randomly chosen images from the Sheffield Dataset was used to train the classifier in each of the 10 runs of the experiments. The training images in the $2\frac{1}{2}$D and composite spaces correspond to the 2D images used in each run of the experiment. The training images were not used for testing. The mean recognition rates along with the error margins (standard deviations) are presented in Table 1.

The accuracy of our baseline recogniser is lower than typical accuracies quoted in the literature. This is because standard benchmark tasks include extensive care in the data collection process, with clear lighting, segmentation and highly restricted pose, and expression variations. Our database has, however, a higher level of variation. In order to confirm the accuracy of our implementations, we tested our single image recogniser on one of the standard benchmark databases, the Yale Database, and were able to reproduce quoted results.

The results of the experiments clearly indicate that our composite image representation that incorporates both 2D intensity information and the $2\frac{1}{2}$D depth information is a significant improvement on either representation by itself. As

expected, increasing the number of training images per class improves the performance of the classifier, with the best results obtained with leave-one-out cross validation using all except 1 of the images in the class for training.

This new representation also addresses the issue of systematic variability in face images as the pose of the subjects varies. The classifier achieves greatest accuracy in the 2D space (62.59%) when the number of training images per class is maximised. Similar accuracy ($61.79\% \pm 1.39$) is obtained in the composite space using only a fraction of the images to train the classifier. This accuracy can be increased further if images of individuals with and without glasses/head-scarf are treated as belonging to the same class, rather than different classes, as is done here.

5. CONCLUSIONS

In this paper we present the composite image representation - a simple yet effective way of combining intensity and depth information for face recognition using Eigenfaces and the Mahalanobis distance. This new representation is tested on the Sheffield Dataset, which is more challenging than many of the publicly available datasets. The composite image representation gives higher recognition rates than the 2D intensity images and the $2\frac{1}{2}$D disparity images, for the same training and test data. It also addresses the issue of systematic variability in the image of face as its pose changes with respect to the camera position. This is highlighted by the fact that this image representation achieves the same accuracy as the 2D representation, but with fewer images, and it is indicative of the value of depth information in automatic face recognition. Furthermore, the use of disparity values instead of actual depth values means that stereo images can be used without the error-prone camera calibration and reconstruction processes. This greatly increases the scope of face recognition applications because the work can, from theoretical considerations, be extended to use dynamic images.

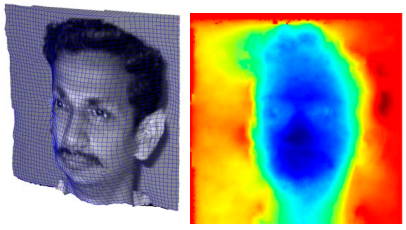

Fig. 1. 3D reconstruction and the disparity map (horizontal displacement) generated from a pair of stereo images (full frontal image shown in Figure 2, image 1). In the disparity map, note high disparities between the image pairs in regions of high depth information, (*e.g. nose*), and low disparity in flat regions (*e.g.* background).

6. REFERENCES

[1] V. Blanz & T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Trans. on PAMI*, vol. 25, no. 9, Sept. 2003.

[2] Chi-Fa Chen & Yu-Shan Tseng & Chia-Yen Chen, "Combination of pca and wavelet transforms for face recognition on 2.5d images," in *Proc. IEEE Conf on Image and Vision Computing*, New Zealand, Nov. 2003, pp. 343–347.

[3] G. Medioni & R. Waupotitsch, "Face modeling and recognition in 3-d," in *Proc. of the IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures (AMFG '03)*, Oct. 2003, pp. 232–233.

[4] I. Yoda & K. Sakaue, "Utilization of stereo disparity and optical flow information for the computer analysis of human interactions," *Machine Vision and Applications*, vol. 13, pp. 185–193, 2003.

[5] N. Uchida & T. Shibahara & T. Aoki & H. Nakajima & K. Kobayashi, "3d face recognition using passive stereo vision," in *Proc. IEEE Int. Conf. on Image Proc. (ICIP '00)*, Sept. 2005, pp. 950–953.

[6] T. Heseltine & N. Pears & J. Austin, "Three-dimensional face recognition using surface space combinations," in *BMVC 2004*, Sept. 2004.

[7] K. Bowyer & K. Chang & P. Flynn, "A survey of 3d and multi-modal 3d+2d face recognition," Tech. Rep. TR 2004-22, University of Notre Dame, France, 2004.

[8] P. J. Phillips & P. J. Flynn & T. Scruggs & K. W. Bowyer & J. Chang & K. Hoffman & J. Marques & J. Min & W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, San Diego, USA, June 2005.

[9] J. Magarey & A. Dick, "Multiresolution stereo image matching using complex wavelets," in *Proc. IEEE Int. Conf. on Pattern Recog. (ICPR)*, Aug. 1998, pp. 4–7.

[10] J. Magarey & N. Kingsbury, "Motion estimation using complex wavelets," *IEEE Tran. on Signal Proc., special issue on Wavelets and Filter Banks*, vol. 46, no. 4, pp. 1069 – 1084, Apr. 1998.

[11] M. Turk & A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71 – 86, 1991.

[12] I. Craw & N. Costen & T. Kato & S. Akamatsu, "How should we represent faces for automatic recognition?," *IEEE Trans. on PAMI*, vol. 21, no. 8, pp. 725 – 736, Aug. 1999.

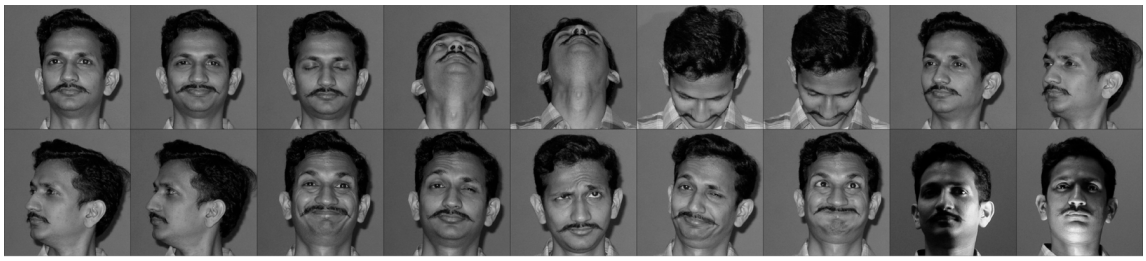

Fig. 2. Sample of images from one of the classes in the Sheffield Dataset.

Method	Training images per class			
	1	5	10	LOO
2D: Left Image Only	33.27 ± 3.15	53.13 ± 2.11	58.58 ± 2.68	**62.59**
$2\frac{1}{2}$D: Disparity Map Only	30.16 ± 2.72	43.95 ± 2.77	46.46 ± 2.85	50.00
Composite Image: 2D + $2\frac{1}{2}$D	37.96 ± 3.82	$\mathbf{61.79 \pm 1.39}$	67.71 ± 2.58	71.48

Table 1. Recognition results for the 2D, $2\frac{1}{2}$D and composite classifiers using sets of 1, 5, 10 and 17 training samples per class. Note the composite classifier is consistently better than the 2D classifier. Further, at only five training examples per face class, it achieves comparable performance to the leave one out classifier which has 17 training examples per class (recognition scores in bold font), demonstrating the usefulness of depth information derived from stereo vision.

Algorithm 1: Image Matching Algorithm

Input: Images A_1, A_2 $(N \times N)$ and levels of decomposition m_{max}
Output: Disparity field SD $(N \times N)$

Perform Complex Discrete Wavelet Transform (CDWT) on A_1 and A_2 using complex valued low-pass and high-pass filters $\{h_0, h_1\}$

$h_0 = [1-j, 4-j, 4+j, 1+j]/10, \quad h_1 = [1-2j, 5+2j, -5+2j, 1-2j]/14$

Output: Six bandpass images $D_1^{(n,m)}$ and $D_2^{(n,m)}$ for A_1 and A_2 at levels $m = 1 : m_{max}$.

for $m = m_{max} : 1$ *// m_{max} is the coarsest level of decomposition*

Compute disparity field at each level using:

for $pix = 1 : p$ *// p: # pixels in level m bandpass images*

Inputs: $\{\mathbf{n}_1, \mathbf{n}_2\}$: (x, y) location of pixels pix to be matched

$\mathbf{f}$: Fractional offsets to give sub-pixel accuracy

$SD^{(m)}(\mathbf{n}_1, \mathbf{n}_2 + \mathbf{f}) = \sum_{n=1}^{6} |D_1^{(n,m)}(\mathbf{n}_1) - D_2^{(n,m)}(\mathbf{n}_2 + \mathbf{f})|^2$ *// sub-band squared difference*

The correspondent for the left pixel $\mathbf{n}_1$ is the subpixel $(\mathbf{n}_2 + \mathbf{f})$ which minimises $SD^{(m)}$

end for

Output: $SD^{(m)}$: Disparity Field at level m

Regularise Disparity Field to eliminate random errors & smooth the field

if $m \neq m_{max}$ *// If this is not the coarsest level*

$SD^{(m)} = SD^{(m)} + SD'^{(m+1)}$

end if

Output: Final disparity field estimate $SD^{(m)}$ at level m

Interpolate & Scale disparity field to account for higher pixel density & decreased spacing between adjacent pixels at level $m - 1$. Propagate interpolated, scaled disparity field to next finer level

end for

Fig. 3. Pseudocode for Magarey's [10] motion estimation algorithm.