Abstract

This paper presents an integrated approach to unconstrained face recognition in arbitrary scenes. The front end of the system comprises of a scale and pose tolerant face detector. Scale normalization is achieved through novel combination of a skin color segmentation and log-polar mapping procedure. Principal component analysis is used with the multi-view approach proposed in [10] to handle the pose variations. For a given color input image, the detector encloses a face in a complex scene within a circular boundary and indicates the position of the nose. Next, for recognition, a radial grid mapping centered on the nose yields a feature vector within the circular boundary. As the width of the color segmented region provides an estimated size for the face, the extracted feature vector is scale normalized by the estimated size. The feature vector is input to a trained neural network classifier for face identification. The system was evaluated using a database of 20 person's faces with varying scale and pose obtained on different complex backgrounds. The face detector was quite robust to all these variations. The performance of the face recognizer was also quite good except for sensitivity to small scale face images. The integrated system achieved average recognition rates of 87% to 92%.

1 Introduction

The difficulties of face recognition lie in the inherent variability arising from face characteristics (age, gender and race), geometry (distance and viewpoint), image quality (resolution, illumination, signal to noise ratio), and image content (background, occlusion, disguise) [14]. Because of such complexity, most face recognition systems [2] assume a well controlled environment and recognize only near frontal faces. However, these constraints need to be relaxed in practice.

Recently, [5, 3, 6] highlighted a framework for unconstrained face recognition in arbitrary scenes: a face detection module to isolate face patterns from complex scenes, followed by a robust recognition scheme to handle varying appearance of faces due to illumination, pose, facial expression and occlusion. We use this framework to propose an integrated face recognition system based on our earlier work [12] as shown in Figure 1. The front end of the system comprises of a scale and pose tolerant face detector based on eigenspaces of log-polar mappings. For a given color input image, the face detector encloses the face from the complex scene within a circular boundary and locates the position of the nose. Radial grid mapping centered on the nose is then performed to extract a feature vector within the boundary. The feature vector is input to a trained neural network classifier for face identification.

Figure 1. An Integrated Face Detection and Recognition System
First, instead of using a multi-resolution pyramid and multiple self-organizing maps for scale invariant face detection, face size is estimated from the skin color segmented image and is used to adjust the spatial extent of the log-polar mapping. This approach is advantageous in terms of computational requirements and performance under large scale variation of faces. Secondly, the log-polar mapping is used for feature extraction. This is used in conjunction with principal component analysis (PCA) and the distance-to-feature space measure suggested in [16, 10] to alleviate the difficulties in obtaining large numbers of non-face patterns [13] for neural network learning.

In [15], the space-variant sampling grids centered on various facial features, extract feature vectors for face recognition. The fixation positions are assumed to be known. In contrast, we used a single scale-normalized uniformly spaced radial grid centered on the nose position detected by the face detector module.

The paper is organized as follows: Section 2 explains the face detection process, and Section 3 explains the face recognition scheme. Section 4 describes the database, and Section 5 provides the experimental results and discussion. Section 6 concludes the paper.

2 Face Detection

Based on the results of our previous work [12] on using radial grid mapping for face recognition, we have extended and modified the standard log-polar mapping procedure to combine with skin color segmentation, and PCA for scale and pose tolerant face detection. For a given color image, face detection is done in 3 stages:

1. Skin color segmentation and morphological operations are performed to find skin blobs. The width of each skin blob is used as an estimate of the face size. The color image is converted to a luminance image using 
   \[ Y = 0.3R + 0.59B + 0.11G \]
   for subsequent processing.

2. For every pixel in a skin blob, a feature vector is derived from log-polar mapping. This mapping grid is centered on the pixel with its largest circle or spatial extent adjusted according to the estimated skin blob width.

3. The log-polar feature vector extracted at every candidate pixel is then projected onto eigenspaces corresponding to different viewpoints and the reconstruction errors are computed. The pixel location at which the global minimum of the reconstruction error is found (over all eigenspaces) is deemed to be the center of the nose on the detected face (because the eigenspaces were constructed from feature vectors obtained by centering the log-polar grid on the nose), if the error is below a threshold.

4. The above 3 steps are repeated for every skin blob in the image.

The details for each stage are elaborated in the following sections.

2.1 Skin Color Segmentation

Skin color segmentation is performed in the chromatic or pure color space using an \textit{a priori} distribution estimated from 20 persons. Pixels with chromatic values \((r,g)\) within the skin color reference range are classified as skin and the rest as background. Morphological operators are then applied to smooth out noise and obtain skin color blobs. The statistics of each skin blob are calculated and used to create a bounding box for it. The width of each blob indicates its size and is used by the log-polar mapping for scale normalization.

2.2 Log-Polar Mapping

Log-polar mapping [1] uses a space-varying sampling grid. The sampling points are located at the intersections of \(N_r\) concentric circles and \(N_\theta\) radial lines. The latter are equally spaced in angle, while the circles are equally spaced in \(r\). The log-polar feature vector is formed by column or row concatenation of the pixels in the log-polar image. The spatial size or width of an object in the image varies according to its distance from the camera. To obtain scale normalized feature vectors, the largest circle in the log-polar grid can be adjusted to enclose the whole object. Figure 2 shows examples of 3 different face sizes from the MIT [16] face database and their scale normalized log-polar images. Observe that despite the face size variations \(^1\) the log-polar face images are similar. In this example, the scales were manually obtained. However, in our actual implementation with color images, the width of the detected skin color blob is used as an estimate of scale.

2.3 Multi-View Face Detection

Given a set of \(m\) log-polar training feature vectors from faces, the basis vectors spanning this space can be obtained from PCA. Dimensionality reduction and compact representation can be achieved by retaining \(k < m\) eigenvectors. A feature vector \(\tilde{\alpha}\) projected onto this low dimensional space and reconstructed as \(\tilde{\alpha}_r\) results in a reconstruction error given by

\(^1\) The face widths vary from 45 to 28 pixels. \(N_r = N_\theta = 32\) was visually found to offer sufficient resolution in the log-polar images.
2.4 Threshold Selection

As mentioned previously, the reconstruction error of a test vector from an eigenspace formed by the log-polar face feature vectors is an indication of the similarity of the test vector to the eigenspace. Hence a threshold can be used on the reconstruction error to accept or reject the test vector as being face representative.

In order to do this, we need to specify acceptable notions of what constitutes true detection, etc. Firstly, we specify a $10 \times 10$ tolerance neighborhood around the center of the nose. Any pixel in this neighborhood can signal a true detection, if it is $i)$ the global minimum of the reconstruction error and $ii)$ if this global minimum is less than the threshold value. A false detection occurs if the above 2 conditions are satisfied outside the $10 \times 10$ 'nose neighborhood'. A false rejection occurs if the global minimum occurs within the nose neighborhood, but is above the threshold. Lastly, a true rejection occurs if the global minimum occurs outside the nose neighborhood, and exceeds the threshold value. The threshold which yields minimum probability of error satisfies

$$p(d_e | \tilde{\alpha} \in \text{face}) P(\tilde{\alpha} \in \text{face}) = p(d_e | \tilde{\alpha} \not\in \text{face}) P(\tilde{\alpha} \not\in \text{face})$$

(2)

where $d_e$ is the reconstruction error and $\tilde{\alpha}$ is a log-polar test vector.

We estimate the above $a$ priori probabilities and $d_e$ distributions using 177 images of people taken in a laboratory environment. For each image, the face is searched within a predefined 800-pixel region. Pixels within the $10 \times 10$ nose neighborhood are considered as face locations, and this yields $P(\text{face}) = \frac{100}{800}$ and $P(\text{Face}) = 1 - P(\text{face})$. $p(d_e | \tilde{\alpha} \in \text{face}) P(\tilde{\alpha} \in \text{face})$ and $p(d_e | \tilde{\alpha} \not\in \text{face}) P(\tilde{\alpha} \not\in \text{face})$ are estimated from the respective histograms of $d_e$, computed over the search regions in all the images. The threshold is finally selected to satisfy equation 2. Some examples of the detected faces are shown in Figure 3 where the scale and pose tolerant nature of face detection are clearly illustrated.

3 Face Recognition

The recognition scheme uses feature vectors generated from a radial grid mapping [12], followed by classification. The feature vectors are generated as follows:

1. Smoothed gradient magnitudes are extracted from the image.

2. A radial grid is overlaid on these features. The grid consists of a set of uniformly spaced concentric circles, divided into segments by radial lines uniformly spaced in angle. The grid is centered on the nose location estimated by the face detection module, and the largest circle is adjusted to the scale provided by the detector.
frontal at medium scale (ms), and frontal at small scale (ss). There were 25 images in each category.

The multi-view face detector developed as described in Section 2 was applied to the database. As mentioned in Section 2, for correct detection, the detected nose location should fall into the $10 \times 10$ neighborhood centered on the nose. The detection rate for all the subjects in plain (scene 1) and complex scenes (scene 2 and 3) are summarized in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>lr</th>
<th>trl</th>
<th>ff</th>
<th>ud</th>
<th>ms</th>
<th>ss</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>scene 1</td>
<td>99.4</td>
<td>99.8</td>
<td>100</td>
<td>98.0</td>
<td>99.8</td>
<td>85.4</td>
<td>97.1</td>
</tr>
<tr>
<td>scene 2</td>
<td>92.8</td>
<td>96.4</td>
<td>98.4</td>
<td>96.0</td>
<td>99.8</td>
<td>86.8</td>
<td>95.0</td>
</tr>
<tr>
<td>scene 3</td>
<td>94.4</td>
<td>98.2</td>
<td>99.4</td>
<td>96.0</td>
<td>98.4</td>
<td>87.4</td>
<td>95.6</td>
</tr>
</tbody>
</table>

Table 1. Face detection or true detection rates on the database described in Section 4. The columns of the table contain classification results (in %) for different category of images in the database. Please refer to text for detailed description.

Table 1 shows that an average face detection rate of between 95% to 97% was achieved for the different scenes. However, it can be observed that the detector performance is somewhat sensitive in two situations. Firstly, the detection rate for images in category lr decreases more noticeably for the complex scenes as compared to other categories. This is because when the head rotates towards the left or right, complex background is included in the detected face as shown in Figure 5. This results in a distorted log-polar feature vector which reduces the detection accuracy.

Secondly, lower detection rates of between 85% to 87% were obtained for the small size face (ss) which is a drop of 11% to 14% from the medium scale face (ms). This gives an idea of the scale tolerance range of the face detector.

To evaluate the integrated face recognition scheme, the face detector results were used for extracting face feature vectors for classification. The parameters used for the radial grid mapping and SOM classifier were considered in [12] where it was found that a radial grid divided into $15 \times 15$ radial and angular segments, and SOM size of $30 \times 30$ were suitable. We used the same values here. As for the RBF network, the input vector was of dimension 225, the hidden layer consisted of 900 nodes (the $30 \times 30$ SOM neurons) and the output layer had 20 nodes.

The training set (1040 images) for the recognizer was formed using only a subset of the plain background images. It comprised of subsets from the lr, trl, ud and ff categories and all these images were at the same scale. The test set was formed from the remaining (1960) images in the plain background as well as all the images from the 2 complex background (6000 images) scenes, with all 3 scales represented.

Having divided the database into two, as above, only the subset of images that were correctly detected were actually used to train and test the recognizer. Here the radial grid face feature vectors were obtained based on the centroid and estimated scale provided by the face detection module. The classification results on the test images using the different classifiers described in Section 3 are shown in Tables 2 for the 3 different backgrounds. The integrated recognition rates were calculated according to the following example: given 100 test images, if the face was correctly detected in 90 of them, these correctly detected faces were passed to the recognition module. Of these, if the recognizer correctly recognized 80 of them, the integrated recognition rate would be reported as 80%.

Among the classifiers evaluated, the RBF consistently out-performed the others, achieving average recognition rates of between 87% to 92%. The slight decrease in recognition rate for face images in lr category is attributed to the inclusion of complex background in the radial grid mapping as illustrated in Figure 5. The recognition rate degrades rapidly from the medium scale (ms) to small scale (ss) face while it is less severe from large scale (ff) to medium scale (ms). In contrast, the detector performs well even at the small scale. The recognition module may be affected by loss of spatial detail in the smaller face or by oversampling by the $15 \times 15$ radial grid (rendering the feature vector sensitive to small image changes). We are investigating this problem further.

To compare with other techniques, [5] reported a recognition rate of between 92.65% to 97.75% on a database of 66 persons with each person having 25 images of different head orientations, illumination condition, and expressions. However, their database does not contain large scale variations seen in our database. Another detection and recognition method [9], based on the Hidden Markov Model reported performance of 86% on the Olivetti Research Ltd database. The ORL database comprises of 400 face images of 40 sub-

Figure 5. The white circle indicates the largest radius of the log-polar mapping while the white cross indicates the centroid position. This figure shows that background is included in the log-polar mapping when the head undergoes left-right rotation.
Figure 3. Examples of detected face despite scale, pose variations and presence of non-face skin regions. The rectangular bounding box indicates non-face skin blob while a white cross on the nose within a circle represents detected face.

3. The mean value of the gradient magnitudes in each segment is used to form the elements of the feature vector.

3.1 Classification Using Self-Organizing Map

The SOM [4] is a sheet like neural network, whose neurons become specifically tuned to various input signal patterns through an unsupervised competitive learning process. To use the trained SOM map for classification, it is calibrated using a majority voting scheme with the labeled training vectors. It is quite possible when the map size is large, or when the number of training vectors is small that some neurons never win a training vector, in which case the neurons are unlabeled. Since a test vector is classified to the nearest neuron’s label, this can lead to a large number of "unrecognized" classifications in the recognition stage. We have experimented with a 2-stage classification scheme, where if a vector was unrecognized using the nearest neuron scheme, it underwent a second attempt at classification using a 5-nearest neuron scheme. In our experiments, SOM1 refers to the nearest neuron classification while SOM2 refers to the 2-stage classification scheme.

3.2 Classification Using Radial Basis Function Neural Network

In addition to SOM classification, we used another classification scheme using the RBF neural network [8]. In our implementation, we used the weights of the neurons in the trained SOM as the centers of the Gaussian radial-basis functions in the hidden layer. The spread of the Gaussian functions is determined based on the minimum Euclidian distance between the SOM neurons. The number of hidden nodes, $m$, is equal to the number of neurons in the SOM. The number of output nodes is equal to the number of persons in our database that we wish to recognize. The weights connecting the hidden nodes to the output nodes are trained such that each output node is tuned to a particular person. During classification, the RBF output node with the highest activation is taken to represent the identity of the image.

4 Database

To evaluate the integrated recognition system, a face database of 20 persons against 2 complex and a plain background were acquired. Examples are shown in Figure 4. These are $130 \times 120$ image sampled from short video sequences. The images include in-and out of plane rotation and scale changes. In all, the database consists of 9,000 images.

Figure 4. Example of three person’s face in the face database with scale, pose and scene variations.

5 Experimental Results and Discussions

For reporting the experimental results, images of each person in the database described in Section 4 were grouped according to pose variations into 6 categories, viz; head rotated from left/right (lr), head tilted (in-plane rotation) left/right (tlr), head up/down (ud), frontal at large scale (ff),
It was found that the system’s performance was slightly sensitive to the inclusion of background in the feature vector when the head undergoes out of plane rotation. The recognition rates were however more sensitive to scale when face size became too small. We are considering methods to improve the performance in this situation.

### References


### Table 2. Integrated face recognition rates (in %) for 20 persons on various scenes. The columns of the table contain classification results for different categories of images in the database. Please refer to text for detailed description of reported recognition rates.

<table>
<thead>
<tr>
<th>Scene</th>
<th>lr</th>
<th>tlr</th>
<th>ft</th>
<th>ud</th>
<th>ms</th>
<th>ss</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNR</td>
<td>99.2</td>
<td>99.8</td>
<td>100</td>
<td>98.0</td>
<td>87.2</td>
<td>51.6</td>
<td>84.3</td>
</tr>
<tr>
<td>SOM1</td>
<td>85.0</td>
<td>90.8</td>
<td>92.9</td>
<td>92.9</td>
<td>49.6</td>
<td>20.4</td>
<td>62.1</td>
</tr>
<tr>
<td>SOM2</td>
<td>95.8</td>
<td>98.3</td>
<td>99.2</td>
<td>98.0</td>
<td>74.4</td>
<td>46.2</td>
<td>74.5</td>
</tr>
<tr>
<td>RBF</td>
<td>99.4</td>
<td>99.8</td>
<td>100</td>
<td>98.0</td>
<td>92.2</td>
<td>57.2</td>
<td>87.1</td>
</tr>
<tr>
<td>Scene</td>
<td>NNR</td>
<td>84.4</td>
<td>96.4</td>
<td>96.6</td>
<td>96.0</td>
<td>81.8</td>
<td>59.4</td>
</tr>
<tr>
<td>SOM1</td>
<td>64.7</td>
<td>95.3</td>
<td>90.3</td>
<td>96.0</td>
<td>52.2</td>
<td>22.7</td>
<td>70.2</td>
</tr>
<tr>
<td>SOM2</td>
<td>81.6</td>
<td>96.4</td>
<td>97.6</td>
<td>96.0</td>
<td>77.4</td>
<td>56.6</td>
<td>84.3</td>
</tr>
<tr>
<td>RBF</td>
<td>92.8</td>
<td>96.4</td>
<td>98.4</td>
<td>96.0</td>
<td>89.8</td>
<td>69.9</td>
<td>90.6</td>
</tr>
<tr>
<td>Scene</td>
<td>NNR</td>
<td>87.8</td>
<td>98.2</td>
<td>99.4</td>
<td>96.0</td>
<td>84.6</td>
<td>59.4</td>
</tr>
<tr>
<td>SOM1</td>
<td>59.3</td>
<td>88.3</td>
<td>81.7</td>
<td>91.3</td>
<td>47.2</td>
<td>23.1</td>
<td>65.2</td>
</tr>
<tr>
<td>SOM2</td>
<td>80.8</td>
<td>98.2</td>
<td>95.4</td>
<td>96.0</td>
<td>73.6</td>
<td>54.0</td>
<td>83.0</td>
</tr>
<tr>
<td>RBF</td>
<td>94.4</td>
<td>98.2</td>
<td>99.4</td>
<td>96.0</td>
<td>90.4</td>
<td>73.6</td>
<td>92.0</td>
</tr>
</tbody>
</table>