Behaviour & Information Technology
Vol. 30, No. 6, November–December 2011, 747–760

Taylor & Francis
Taylor & Francis Group

Video-based person recognition using fovea intensity comparison code

M. Balasubramanian*, S. Palanivel and V. Ramalingam

Department of Computer Science and Engineering, Annamalai University, Annamalainagar 608 002, India

(*Received 28 April 2009; final version received 16 September 2009*)

This article proposes a feature extraction method for automatic person recognition in video. The method proposed by Viola and Jones (Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: *IEEE international conference on computer vision and pattern recognition (CVPR 2001)*, Kauai Marriott, Hawaii, Vol. 1, 511–518) is used to detect the face region. Face region is processed in YC_bC_r colour space to determine the locations of the eyes. The centre of the mouth is determined relative to the locations of the eyes. Facial and mouth features are extracted using multiscale morphological erosion and dilation operations, respectively. The facial features are extracted relative to the locations of the eyes, and mouth features are extracted relative to the locations of the eyes and mouth. Fovea intensity comparison code and exclusive-OR operation for matching are used to recognise a person in video sequences. Projected fovea intensity comparison code (PFICC) and Euclidean distance for matching are also used to recognise a person in video sequences. The performance of the system using PFICC is evaluated in real time in the laboratory environment, and the system achieves a recognition rate (RR) of 99.0% and an equal error rate (EER) of about 0.84% for 50 subjects. The performance of the system is also evaluated for the eXtended Multi Modal Verification for Teleservices and Security (XM2VTS) database, and the system achieves an RR of 100% and an EER of about 0.23%.

Keyword: face detection; eye location; multiscale morphological dilation operation; multiscale morphological erosion operation; fovea intensity comparison code

1. Introduction

Automatic face recognition by machine appears to be difficult, while it is done effortlessly by human beings. The main reason for this difficulty is that it is difficult to articulate the mechanism used by humans. Face recognition can be categorised into face identification and authentication. The objective of a face identification system is to determine the identity of a test subject from the set of reference subjects. The performance of the face identification system is quantified in terms of identification rate or recognition rate (RR). On the other hand, a face authentication system should accept or reject the identity claim of a subject, and the performance is measured in terms of equal error rate (EER). Person authentication systems make use of one or more biometric modalities such as speech, face, fingerprint, signature, iris and hand geometry to accept or reject the identity claim of an individual. In this article, face and mouth modalities are used for person recognition. The terms facial and mouth features refer to the features extracted from the face and mouth image of the person, respectively.

1.1. Related work

Several techniques have been proposed in the literature for still-image-based face recognition such as principal component analysis (PCA) or eigenface analysis (Turk and Pentland 1991), linear discriminant analysis (Swets and Weng 1996, Belhumeur *et al*. 1997), independent component analysis (Bartlett *et al*. 1998), elastic graph matching (Lades *et al*. 1993, Wiskott *et al*. 1997), line edge map (Gao and Leung 2002), support vector machine (Kotropoulos and Pitas 2001, Heisele *et al*. 2002) and correlation filter (Savvides *et al*. 2002). Most of the video-based face recognition methods apply still-image-based recognition to selected frames (Zhao *et al*. 2000). The radial basis function neural network (Howell and Buxton 1996), probabilistic modelling (Zhou *et al*. 2003) and hidden Markov model (HMM) (Liu and Chen 2003) are also used for video-based face recognition. A comprehensive survey of automated face recognition techniques can be found in Zhao *et al*. (2003). The problem of face recognition over changes in illumination is widely recognised to be difficult for humans (O'Toole *et al*. 2007). A fully automatic face recognition algorithm and its performance on the

*Corresponding author. Email: balu_june1@yahoo.co.in

ISSN 0144-929X print/ISSN 1362-3001 online

http://dx.doi.org/10.1080/01449290903353047
http://www.tandfonline.com

FRGC v2.0 data were presented by Mian *et al*. (2007). The computational tools and a hardware prototype for three-dimensional face recognition were presented by Kakadiaris *et al*. (2007). The mouth features such as discrete cosine transform of the lip region (Chaudhari *et al*. 2003), eigenlips (Jourlin *et al*. 1997, Kanak *et al*. 2003) are commonly used to represent the mouth image. The video-based face recognition system called PersonSpotter described by Steffens *et al*. (1998) used the elastic graph matching technique. An RR of about 90.0% was reported. A method proposed by Zhou *et al*. (2003) used probabilistic modelling of intensity values of the images, and a recognition performance of about 98.0% was reported using the MoBo database (Gross and Shi 2001) having 25 subjects. The method described by Liu and Chen (2003) used PCA and HMM, and it reported 98.8% RR for the MoBo database. Facial image representation based on local binary pattern texture features is given by Timo Ahonen and Inen (2006). The logarithmic total variation model for face recognition under varying illumination, including natural lighting conditions was described by Terrence *et al*. (2006). The average face RR of 99.65% was reported for Carnegie Mellon University (CMU) Pose, Illumination and Expression (PIE) database (Sim *et al*. 2002). Xiang *et al*. (2006) used recursive Fisher linear discriminant for face recognition and an RR of 99.65% was reported for CMU PIE database.

1.2. Contributions and outline of work

Most of the existing person recognition methods use raw pixel values of the face image for recognition. Features usually encode knowledge about domain and it is difficult to learn from raw pixel values. In this work, local minima (maxima) features are extracted from the facial (mouth) regions (fovea regions) and the extracted features are compared to form the fovea intensity comparison code (FICC). The facial features such as hair, face outline, eyebrows, eyes and mouth play an important role in perceiving and remembering faces. A cartoonist extracts the required information from these features and represents in terms of lines and arcs. These lines and arcs correspond to the gradient or local extrema (minima and maxima) in an image. The local maxima and minima are the largest and smallest intensity values of an image within some local neighbourhood, respectively.

The automatic person recognition system described in this article consists of four modules: face detection, facial and mouth feature extraction, FICC and matching. Face detection is described in Section 2. Facial and mouth feature extraction are described in Sections 3 and 4, respectively. Section 5 describes FICC and matching. Experimental results are given in Section 6. Section 7 concludes the article.

2. Face detection

Detecting faces automatically from the intensity or colour image is an essential task for many applications like person authentication and video indexing. A number of techniques have been proposed for face tracking and localisation (Rowley *et al*. 1998, Fleuret and Geman 2001, Yang *et al*. 2002), we used the method proposed by Viola and Jones (2001) for determining the face region in the video. In the first step, simple rectangle features as shown in Figure 1 are used. These features are reminiscent of Haar basis functions. The rectangle features are computed quickly using an intermediate representation by the sum of the pixels above and to the left of x, y inclusive:

$$\mathrm{II}(x, y) = \sum_{x' \le x,\, y' \le y} \mathrm{I}(x', y') \tag{1}$$

where II(x, y) is the integral image and I(x, y) is the original image. The set of rectangle features used provide rich image representation, which supports effective learning. In the second step, Adaboost learning algorithm is used to select a small set of features and to train the classifier.

About 180,000 rectangle features are associated with each image's sub-window. Out of these large numbers of features, a very small number of features are combined to form an effective classifier. In the third step, a cascade of classifiers as shown in Figure 2 is constructed, which increases the detection performance thereby reducing the computation time.

A positive result from the first classifier triggers the evaluation of the second classifier which triggers the

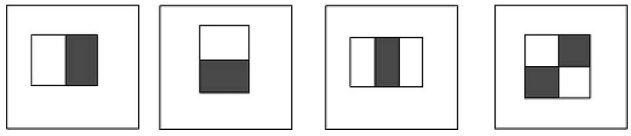

Figure 1. Four types of Harr-like features.

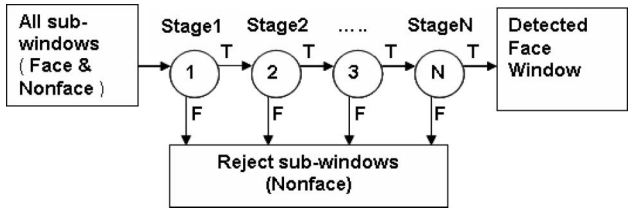

Figure 2. Detection cascade.

third classifier and so on. A negative result from any stage is rejected. The stages in the cascade are constructed using Adaboost algorithm and the threshold is adjusted to minimise false negatives. Stages are added until the target for false positive and detection rate is achieved. A rectangle is fitted over the detected face region. The detected face region is shown in Figure 3.

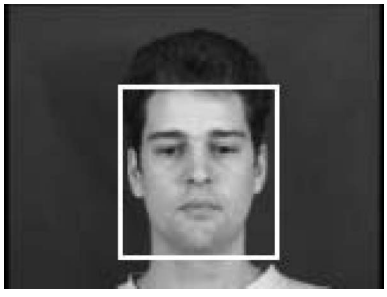

Figure 3. Detected face region.

3. Facial feature extraction

One of the main issues in constructing an automatic person recognition system is to extract the facial or mouth features that are invariant to the size of the face. The face tracking/localisation method gives only the upright rectangular bounding box for the face region, and hence the size of the face cannot be determined from the bounding box. The size of the face can be determined if the locations of two or more facial features are identified. Among the facial features, eyes and mouth are the most prominent features used for determining the size and pose of the face (Nikolaidis and Pitas 2000, Hsu *et al*. 2002). In this section, a method is proposed for extracting the facial features from the face. The facial features are extracted relative to the locations of the eyes.

3.1. Determination of eye location

Several techniques have been proposed in the literature for determining the locations of the eyes. The template-based approach is commonly used for locating the eyes (Lam and Yan 1996, Smeralsi *et al*. 2000), and the methods given by Nikolaidis and Pitas (2000) and Hsu *et al*. (2002) use the gray-scale morphological operations, dilation and erosion (Jackway and Deriche 1996). Nikolaidis and Pitas (2000) applied the morphological operations on the image to enhance the dark regions, and Hsu *et al*. (2002) used morphological operations to emphasise brighter and darker pixels in the luminance (Y) component around the eye regions. In addition to the luminance component, the blue and red chrominance (C_b and C_r) information is also used by Hsu *et al*. (2002). The YC_bC_r colour space is obtained from RGB colour space using

$$\begin{cases} Y = & 0.299R + 0.587G + 0.114B \\ C_b = & B - Y \\ C_r = & R - Y \end{cases} \tag{2}$$

where R, G and B are the red, green and blue components of the colour image, respectively. The RGB and YC_bC_r representation of the face region are shown in Figure 4.

The Y, C_b and C_r values are normalised to the range [0, 255]. The eye regions have low intensity (Y), high blue chrominance (C_b) and low red chrominance (C_r) when compared to the forehead region of the face. Using this fact, the face region is thresholded to obtain the thresholded face image U, given by

$$U(i,j) = \begin{cases} 255, & \text{if } Y(i,j) < \lambda_1 \text{ and } C_b(i,j) > \lambda_2 \\ & \text{and } C_r(i,j) < \lambda_3 \\ I(i,j), & \text{otherwise} \end{cases} \tag{3}$$

where λ_1, λ_2 and λ_3 are the average Y, C_b and C_r values of the pixels in the forehead region, respectively. The forehead region is determined from w_1 and (c_x,c_y). Figure 5 shows the construction of the thresholded face image. The white blobs in Figure 5(a)–(c) are respectively the low intensity, high blue chrominance and low red chrominance regions when compared to the forehead region of the face. The threshold face image is shown in Figure 5(d). Morphological closing operation is applied to the thresholded face image, and the centroids of the blobs are determined.

The relative positions of the centroids with respect to the rectangular bounding box enclosing the face region and the centre (average) of the eyebrow pixel coordinates are used to determine the locations of the eyes. The eyebrow (E) pixel coordinates are obtained based on the change in the gray values in the eyebrow region.

$$E(i,j) = \begin{cases} 1, & \text{if } Y(i,j) \geq \lambda_1 \text{ and } Y(i,j+1) \geq \lambda_1 \\ & \text{and } Y(i,j+2) < \lambda_1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Figure 6a shows the centroids of the white blobs in the thresholded face image, and Figure 6b shows the locations of the eyes.

3.2. Facial feature extraction

Facial feature extraction is an interesting and challenging task. The key facial features such as hair,

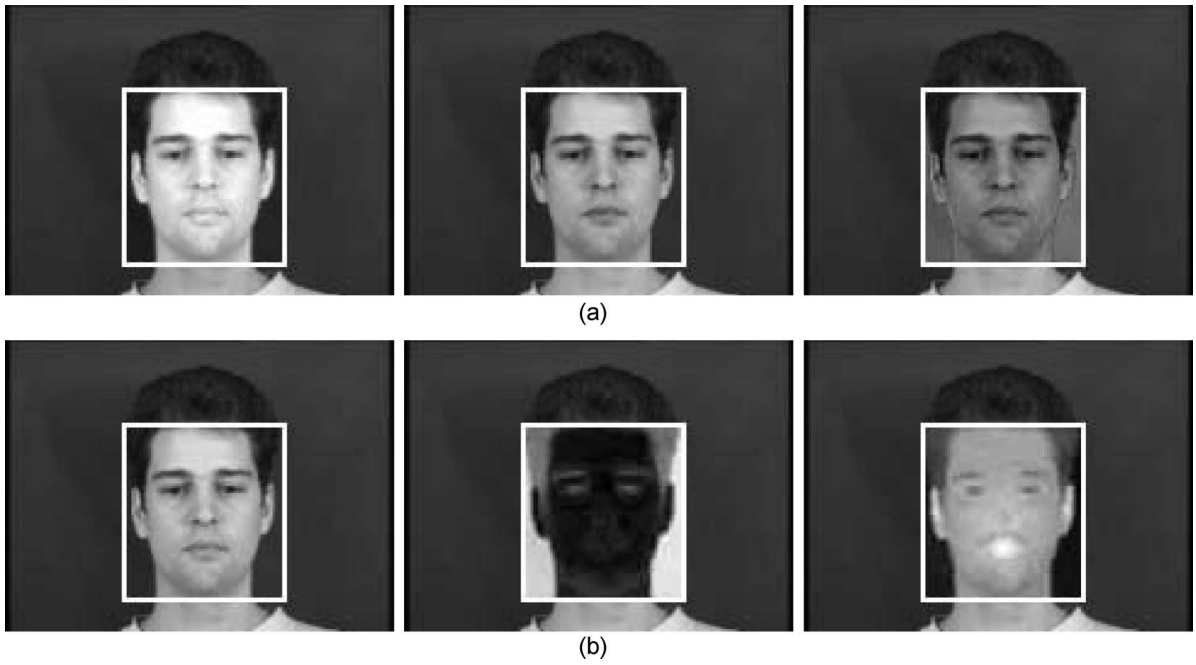

Figure 4. *RGB* and YC_bC_r representation of the face region. (a) From left to right: red, green and blue components of the face region (b) Y,C_b and C_r components of the face region.

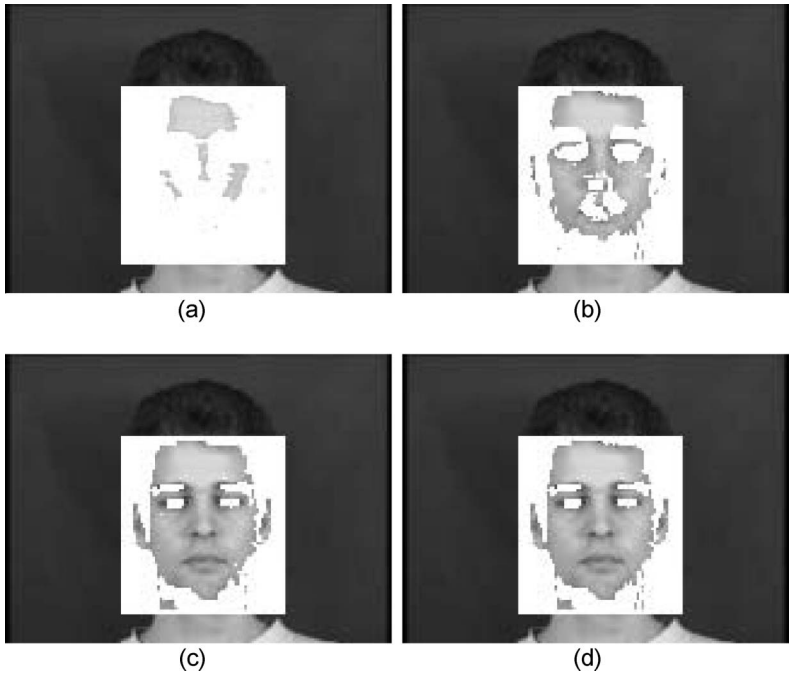

Figure 5. Construction of thresholded face image. The white blobs in (a), (b) and (c) are respectively the low intensity, high blue chrominance and low red chrominance regions when compared to the forehead region of the face. (d) Thresholded face image.

eyebrows, eyes, nostrils and end points of lips are associated with local minima, and the shape of the lip contour and nose tip corresponds to local maxima. The local maxima and minima can be extracted using the gray scale morphological operations such as dilation and erosion, respectively (Jackway and Deriche 1996). The morphological dynamic link architecture method for face recognition described by Kotropoulos *et al.*

Figure 6. Determination of eye location. (a) Centroids of the blobs. (b) Locations of the eyes.

(2000) uses multiscale morphological dilation and erosion under the elastic graph matching frame work. In our method, an elliptical rigid grid is placed over the face region, and the multiscale morphological erosion is used for feature extraction. Most of the key facial features are associated with the local minima, and hence we use only the erosion operation for facial feature extraction. The elliptical grid is used instead of a rectangular grid (Kotropoulos *et al.* 2000) in order to extract features only from the face region. In the elliptical grid, the grid points (nodes) lie within an elliptical region. The grid points are selected in order to extract local features within some local neighbourhood of each grid point. The grid points are placed non-uniformly over the face image based on experimental studies. The face outline or contour can be captured using a rectangular grid which assumes that the training and testing images have the same background. The performance of the person recognition technique must be invariant to the position of the face in the image, and hence we use an elliptical grid instead of a rectangular grid. The length and the slope of the line connecting the eyes are used to determine the size and orientation of the grid, respectively. The elliptical grid consists of 73 nodes, and the positions of these nodes are determined relative to the locations of the eyes. The multiscale morphological erosion operation is applied at each grid node for extracting the facial features as described below.

The multiscale morphological erosion operation is based on the gray scale morphology, erosion. Let $\mathbb{Z}$ denote the set of integer numbers. Given an image I: $D \subseteq \mathbb{Z}^2 \rightarrow \mathbb{Z}$ and a structuring function G_σ: $G_\sigma \subseteq \mathbb{Z}^2 \rightarrow \mathbb{Z}$ at scale σ, the erosion of the image I by the structuring function G_σ is denoted as $(I \ominus G_\sigma)$, and it is defined by

$$(I \ominus G_\sigma)(i,j) = \min_{x,y} \{I(i+x, j+y) - G_\sigma(x,y)\} \quad (5)$$

where $-m_a \leq x,y \leq m_b$, with $1 \leq i \leq w$, $1 \leq j \leq h$. The size of the structuring function is decided by the parameters m_a and m_b, and is given by $(m_a + m_b + 1) \times (m_a + m_b + 1)$. The structuring functions such as flat, hemisphere, paraboloid are commonly used in morphological operations (Jackway and Deriche 1996). The flat structuring function $G_\sigma(x, y) = 0$ is used in this article. For a flat structuring function, the expression for erosion reduces to

$$(I \ominus G_\sigma)(i,j) = \min_{x,y} \{I(i+x, j+y)\} \quad (6)$$

where $-m_a \leq x,y \leq m_b$. The erosion operation (6) is applied at each grid node for $\sigma = 1,2, \ldots, p$ to obtain p facial feature vectors from the face image. The distance between the eyes (d_e) is used to determine the parameters m_a, m_b and p. The value $m_a = \lfloor d_e/32 \rfloor + \lfloor \sigma/2 \rfloor$, $m_b = \lfloor d_e/32 + 0.5 \rfloor + \lfloor (\sigma-1)/2 \rfloor$ and $p = 3$ has been used in our experiments. These parameters are chosen in such a way that $m_a + m_b + 1$ for $\sigma = p$ is less than or equal to the minimal distance between two nodes of the grid which depends on the number of nodes in the grid. Figure 7(a) shows the eroded images for $\sigma = 1,2$ and 3. Figure 7(b) shows the facial regions used for extracting the feature vectors for $\sigma = 1,2$ and 3.

4. Mouth feature extraction

One of the main issues in constructing an automatic person authentication system is to extract the mouth features that are invariant to the size of the face and mouth. In this section, a method is proposed for extracting mouth features from the mouth image. The mouth features are extracted relative to the locations of the eyes and mouth.

4.1. Determination of mouth centre

The mouth or lip image analysis has received considerable attention in the area of speech recognition and person recognition. Mouth image segmentation is a necessary step for real time mouth feature extraction.

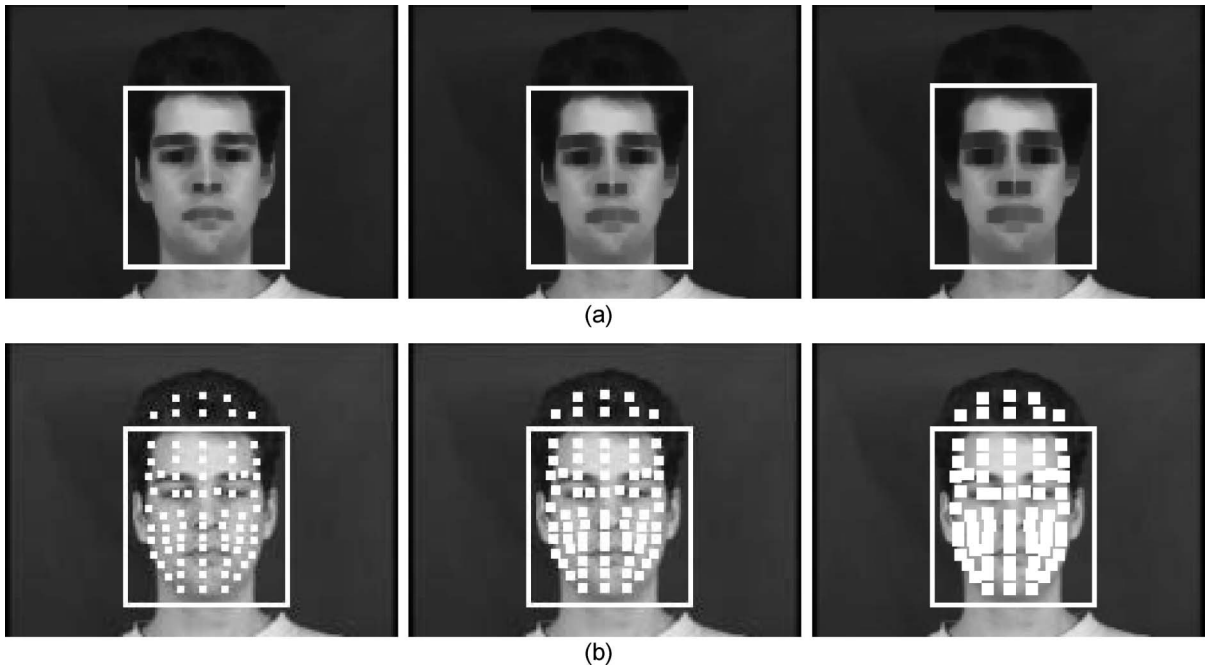

Figure 7. Facial feature extraction. (a) From left to right: eroded images for $\sigma = 1$, 2 and 3. (b) Facial regions used for extracting the feature vectors for $\sigma = 1$, 2 and 3.

Recent methods (Chen 2001, Hsu *et al*. 2002, Leung *et al*. 2004, Wang *et al*. 2004) use colour information to distinguish the lip and nonlip regions in the face. The term lip image or lip region refers to the lips, teeth, moustache and the interior of the mouth. For face images with weak colour contrast, accurate and automatic extraction of inner and outer lip boundaries remains a challenging task. Different types of facial hair in the mouth region complicates the lip contour extraction or the lip contour itself may not be visible. The mouth region is determined from the locations of the eyes and the centre of the mouth. The centre of mouth is calculated using

$$m_{cx} = e_{cx} + 1.1d\cos(\theta + 1.57) \tag{7}$$

$$m_{cy} = e_{cy} + 1.1d\sin(\theta + 1.57) \tag{8}$$

$$\theta = \tan^{-1}\left(\frac{e_{ry} - e_{ly}}{e_{rx} - e_{lx}}\right) \tag{9}$$

where m_{cx} and m_{cy} are x and y-coordinates of the mouth centre, respectively. e_{cx} and e_{cy} are the average x and y-coordinates of the locations of the eyes, respectively. e_{lx} and e_{ly} are x and y-coordinates of the location of left eye. e_{rx} and e_{ry} are x and y-coordinates of the location of right eye. d is the distance between the eyes and θ is the slope of the line connecting the locations of the eyes with respect to x-axis. The centre of the mouth is shown in Figure 8.

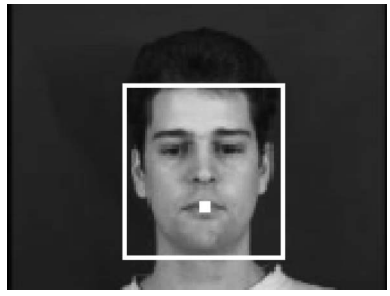

Figure 8. Centre of the mouth.

4.2. *Mouth feature extraction*

The static nature of the mouth or appearance of the mouth image over a period of time characterises an individual to some extent. The shape of the lip contour and shape of the moustache are the dominant mouth features in the mouth region. These features are associated with local maxima because the lip, moustache and the interior of the mouth have low luminance (Y) than the nonlip region. The local maxima can be extracted using the morphological dilation (Jackway and Deriche 1996). For real time mouth feature extraction from the mouth image, a

rectangular grid consisting of 25 nodes is placed over the mouth region. The positions of these nodes are determined relative to the locations of the eyes and mouth. The features are extracted at each grid node using the multiscale morphological dilation operation as described below.

Given an image I: $D \subseteq \mathbb{Z}^2 \rightarrow \mathbb{Z}$ and a structuring function G_σ: $G_\sigma \subseteq \mathbb{Z}^2 \rightarrow \mathbb{Z}$ at scale σ, the dilation of the image I by the structuring function G_σ is denoted as $(I \oplus G_\sigma)$, and it is defined by

$$(I \oplus G_\sigma)(i,j) = \min_{x,y} \{I(i-x, j-y) + G_\sigma(x,y)\} \quad (10)$$

where $-m_a \leq x,\ y \leq m_b$, with $1 \leq i \leq w,\ 1 \leq j \leq h$. For a flat structuring function, the dilation can be expressed as

$$(I \oplus G_\sigma)(i,j) = \min_{x,y} \{I(i-x, j-y)\} \quad (11)$$

where $-m_a \leq x,y \leq m_b$. The dilation operation (11) is applied at each grid node for $\sigma = 1,2, \ldots, p$ to obtain p mouth feature vectors from the mouth image. The distance between the eyes (d_e) is used to determine the parameters m_a, m_b and p. The value $m_a = \lfloor d_e/64 + 0.5 \rfloor + \lfloor (\sigma - 1)/2 \rfloor$, $m_b = \lfloor d_e/64 \rfloor + \lfloor \sigma/2 \rfloor$ and $p = 3$ has been used in our experiments. Figure 9(a) shows the dilated images for $\sigma = 1,2$ and 3. Figure 9(b) shows the visual regions used for extracting the feature vectors for $\sigma = 1$, 2 and 3.

5. Fovea intensity comparison code and matching

5.1. *Fovea intensity comparison code*

The FICC is obtained from the m-dimensional feature vector ($\boldsymbol{x}$) extracted from the face (mouth) image using

$$y[m(i-1) + j - 0.5i(i+1)] = \begin{cases} 1, & \text{if} x(i) < x(j) \\ 0, & \text{otherwise} \end{cases} \quad 1 \leq i \leq m, 1 \leq j \leq m, \text{ and } i < j \quad (12)$$

y is the d-dimensional FICC, where $d = (m^2 - m)/2$. In Equation (12), the intensity of each fovea region is compared with the intensity of other fovea regions to form the FICC.

The 2628-dimensional face FICC is obtained from the 73-dimensional feature vector extracted from the face image using Equation (12). Similarly, 300-dimensional mouth FICC is obtained from the 25-dimensional feature vector extracted from the mouth image.

5.2. *Matching of fovea intensity comparison codes*

The similarity between FICC is calculated using exclusive-OR operation. The d-dimensional FICC for

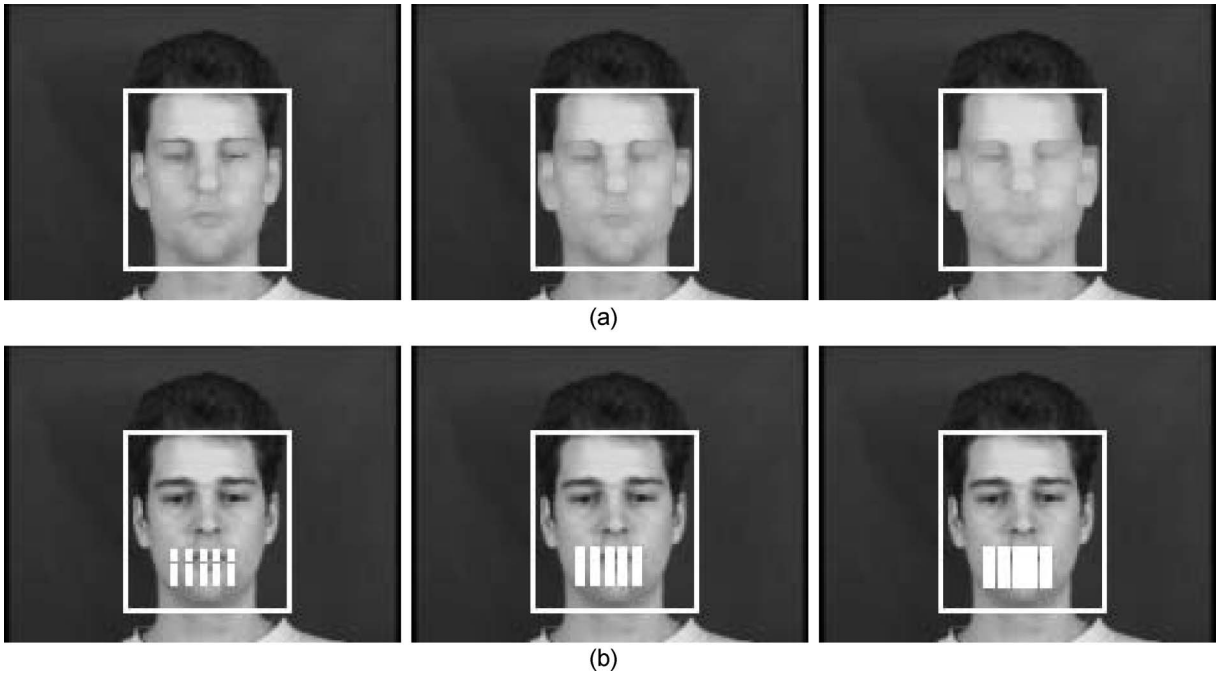

Figure 9. Mouth feature extraction. (a) From left to right: dilated images for $\sigma = 1$, 2 and 3. (b) Mouth regions used for extracting the feature vectors for $\sigma = 1$, 2 and 3.

a test face (mouth) image is compared with the FICC of training face (mouth) images to obtain the error for each subject $e(k)$, using

$$e(k) = \frac{1}{d}\sum_{j=1}^{j=d} y_{\text{test}}(j) \oplus y_{\text{train}}(k)(j), 1 \leq k \leq n \qquad (13)$$

where n is the number of subjects, $s(k)$ is the error for kth subject, y_{test} and y_{train} are the feature vectors extracted from test image and train image, respectively, and $\oplus$ is an exclusive-OR operation.

The 2628-dimensional FICC for a test face image is compared with the FICC of training face images to obtain the error for each subject using Equation (13). Similarly, 300-dimensional FICC for a test mouth image is compared with the FICC of training mouth images to obtain the error for each subject. The identity of the test image is decided based on the lowest error.

5.3. *Projected fovea intensity comparison codes and matching*

PCA (Turk and Pentland 1991) is a multivariate statistical technique for reducing the dimension of a vector (d) to a lower dimension (p) using orthogonal factor space. Then, by minimum distance matching, the projected test vector ($\mathbf{t}$) can be assigned to the class (c) corresponding to the projected training vector $\boldsymbol{g}_i$, where

$$c = \arg\min_{1 \leq i \leq n} ||\mathbf{t} - \mathbf{g}_i|| \qquad (14)$$

where $||\cdot||$ represents the Euclidean distance in $\mathbb{R}^p$.

The 2628-dimensional FICC of face images are projected and reduced into 30-dimensional vectors using PCA. The 30-dimensional projected FICC of a test face image is compared with the projected FICC of training face images to obtain the identity of the test face image using Equation (14). Similarly, 300-dimensional FICC of mouth images are projected and reduced into 20-dimensional vectors. The 20-dimensional projected FICC of a test mouth image is compared with the projected FICC of training mouth images to obtain the identity of the test mouth image. The projected face feature vectors for male and female subjects are shown in Figures 10 and 11, respectively. The projected mouth feature vectors for male and female subjects are shown in Figures 12 and 13, respectively.

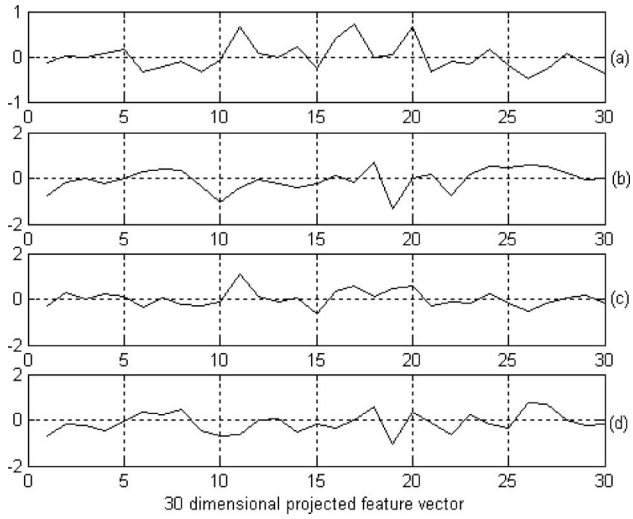

Figure 10. Projected face feature vectors for male subjects: (a) and (c) subject 1; (b) and (d) subject 2.

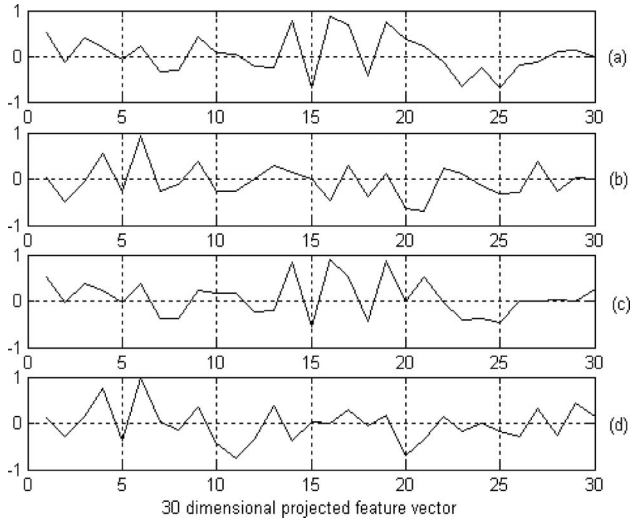

Figure 11. Projected face feature vectors for female subjects: (a) and (c) subject 1; (b) and (d) subject 2.

6. Experimental results

The performance of person recognition using FICC is evaluated using XM2VTS database and in the laboratory environment for 50 subjects using a camera with a resolution of 160 × 120. Two hundred and forty-eight subjects of the XM2VTS audio-visual database (Messer *et al.* 1999) are used for our experiments. The XM2VTS database consists of video data recorded from 295 subjects in four sessions, spaced monthly. The first recording per session of the phonetically balanced sentence ('Joe took father's green shoe bench out') is used. The original frame resolution of 720 × 576 is downsampled to 320 × 240, and is used. For person recognition, we used the method described by Viola and Jones (2001) for determining the face region as described in

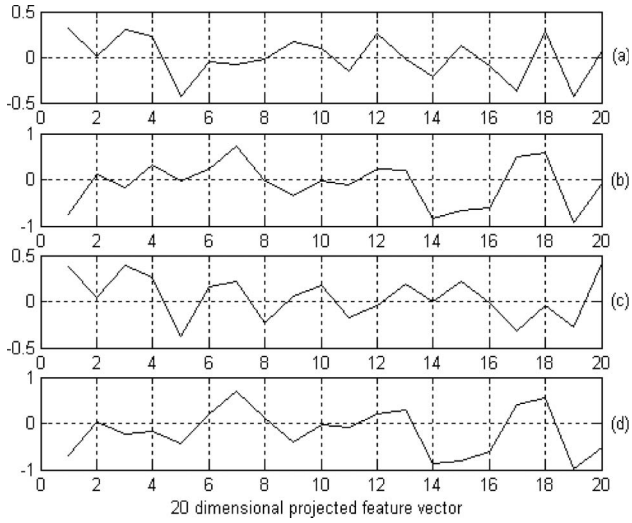

Figure 12. Projected mouth feature vectors for male subjects: (a) and (c) subject 1; (b) and (d) subject 2.

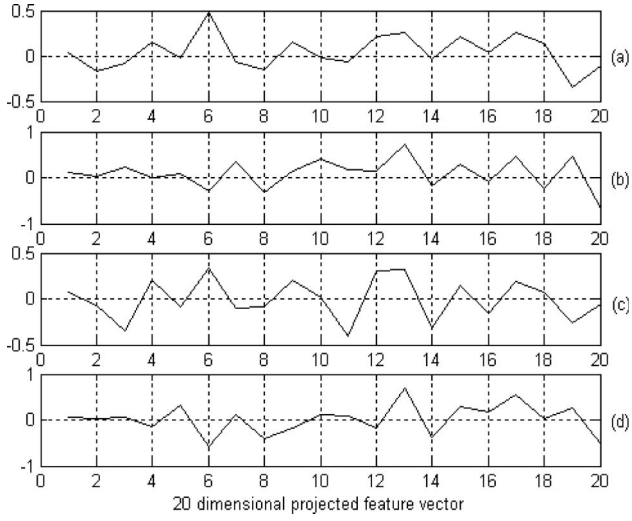

Figure 13. Projected mouth feature vectors for female subjects: (a) and (c) subject 1; (b) and (d) subject 2.

Section 2. The locations of the eyes and mouth are determined as described in Sections 3 and 4, respectively. The method can detect the locations of the eyes in the presence of eye glasses as long as the eye regions are visible. The experimental results show that the method correctly detects the locations of eyes about 95% of the frames in the video. The facial and mouth features are extracted only if all the 73 nodes lie with in the face/head region. Figure 14 shows the face region, locations of the eyes and mouth in real time for a few subjects.

For enrolling a subject, facial and mouth features are extracted in real time as described in Sections 3 and 4 for 50 face (mouth) images with variations in size, orientation, appearance, expressions and pose of the face. The morphological erosion (dilation) is applied on the face (mouth) image for three different scales ($p = 3$). The distance between the eyes varied from 24 to 33 pixels and hence the value of $p = 3$ is used in our experiments.

In training phase, 73-dimensional feature vector is extracted from face image and is converted into 2628-dimensional FICC as described in Section 5 for three different scales. For training, 7500 face feature vectors are extracted from 50 subjects (150 feature vectors per subject). Similarly, 25-dimensional feature vector is extracted from mouth image and is converted into 300-dimensional FICC as described in Section 5 for three different scales. For training, 7500 feature vectors are extracted from 50 subjects.

For identification, 73-dimensional face feature vector is extracted in real time, and is converted into 2628-dimensional FICC as described in Section 5 for three different scales. The 2628-dimensional FICC for a test mouth image is compared with the FICC of training mouth images to obtain the error for each subject using Equation (13) for three different scales. Similarly, 25-dimensional mouth feature vector is extracted in real time, and is converted into 300-dimensional FICC as described in Section 5 for three different scales. The 300-dimensional FICC for a test mouth image is compared with the FICC of training mouth images to obtain the error for each subject using Equation (13) for three different scales. The identity of the test image is decided based on the lowest error. The identification performance is measured in terms of RR.

For authentication, 73-dimensional face feature vector is extracted in real time, and is converted into 2628-dimensional FICC as described in Section 5 for three different scales. The 2628-dimensional FICC for a test mouth image is compared with the FICC of training mouth images of the respective subject using Equation (13) for three different scales. Similarly, 25-dimensional mouth feature vector is extracted in real time, and is converted into 300-dimensional FICC as described in Section 5 for three different scales. The 300-dimensional FICC for a test mouth image is compared with the FICC of training mouth images of the respective subject using Equation (13) for three different scales. The claim is accepted if the error is less than a threshold, otherwise the claim is rejected. The identification and authentication experiments are repeated for one more time for each subject. Similar experiments are carried out using projected face and mouth feature vectors for identification and authentication.

In the database of 50 subjects, there are 50 authentic claims and 49×50 impostor claims.

The authentication performance is measured in terms of EER, where the false acceptance rate and false rejection rate are equal. The EER can be found for each subject (person-specific threshold) or considering all the subjects together (person-independent threshold). In our experiments, the EER is obtained by employing person-independent thresholds.

Similar experiments are carried out for the XM2VTS database (Messer *et al*. 1999). The identification and authentication performance of the system is evaluated for the single and combined modalities in real time and XM2VTS database using the proposed method and the existing methods such as PCA, FLD and RFLD. Performance of the system for single and combined modalities is given in Table 1. The output from face and mouth modalities are combined using Equation (15) to obtain the multimodal or combined error ($e_c(k)$).

$$e_c(k) = w_f e_f(k) + w_m e_m(k), 1 \leq k \leq n \qquad (15)$$

where $e_f(k)$ and $e_m(k)$ are the facial and mouth error for kth subject, respectively. The weight for each of the modality is decided by the parameters w_f and w_m. Equation (15) is applied for each frame in the test video and average error is used for identification and authentication. In our experiments, the modalities are combined using $w_f = 0.7$ and $w_m = 0.3$. The values of the parameters w_f and w_m are chosen such that the system gives optimal performance in terms of EER for the combined modality.

Figures 15 and 16 show the real time facial and mouth feature extraction of varying size, orientation

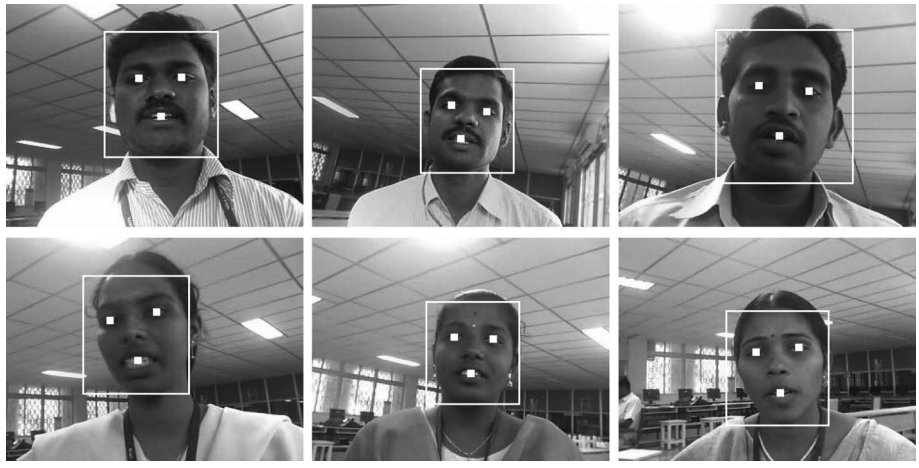

Figure 14. Face region, locations of the eyes and mouth of a few subjects.

Table 1. Person recognition results.

	Real time			XM2VTS database		
	Face (%)	Mouth (%)	Face + mouth (%)	Face (%)	Mouth (%)	Face + mouth (%)
FICC						
RR	96	94	97	98	96	100
EER	1.14	1.64	0.89	0.41	0.47	0.27
Projected FICC						
RR	98	94	99	98	96	100
EER	1.04	1.52	0.84	0.37	0.42	0.23
PCA						
RR	92	90	95	94.5	92	96
EER	1.37	1.75	1.12	0.56	0.62	0.42
FLD						
RR	94	92	96	96.75	94	98
EER	1.3	1.68	1.04	0.52	0.62	0.38
RFLD						
RR	95	92	96	98	94	99
EER	1.26	1.68	1.02	0.49	0.62	0.35

and background, and feature extraction with variations in mouth appearance and expressions are shown in Figure 17. Figures 18 and 19 show the facial and mouth feature extraction using XM2VTS database. Figure 20 shows the snapshot of the video-based person recognition system using FICC. The person recognition system tracks the face, determines the locations of the eyes, extracts mouth features and calculates the output in real time at about 12 frames/s on a personal computer with 3.06 GHz central

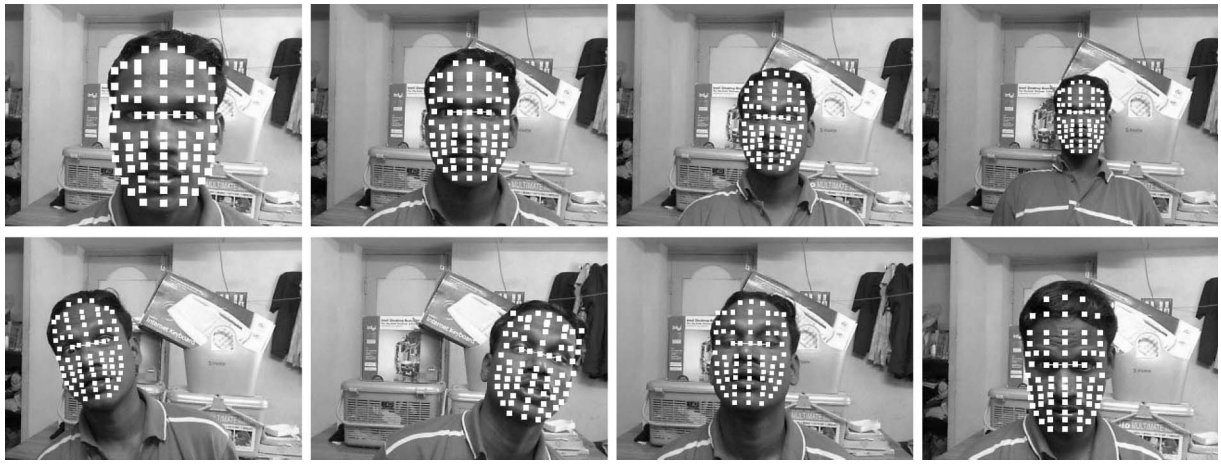

Figure 15. Real time facial feature extractions of varying sizes, orientations and backgrounds.

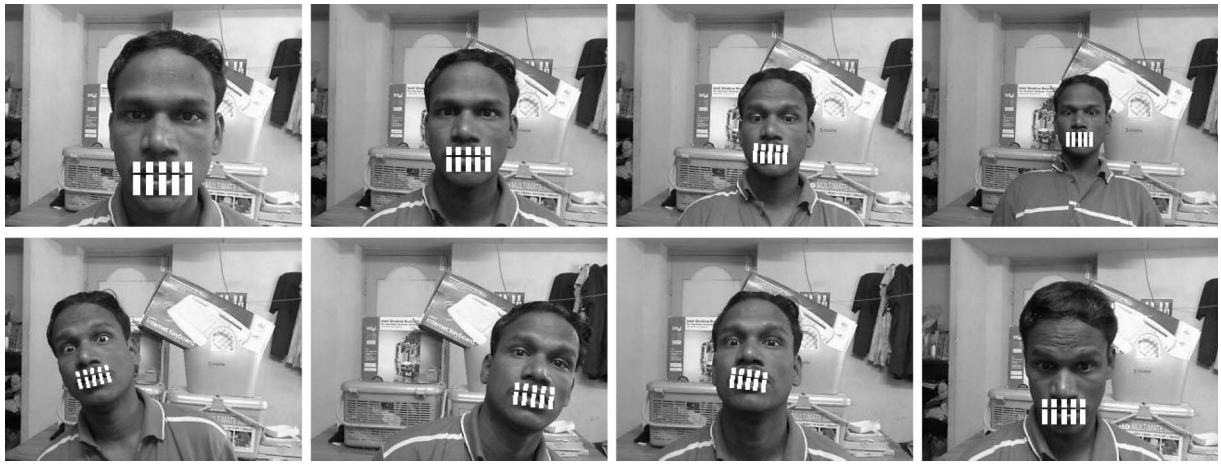

Figure 16. Real time mouth feature extractions of varying sizes, orientations and backgrounds.

Figure 17. Real time mouth feature extractions of varying appearances and expressions.

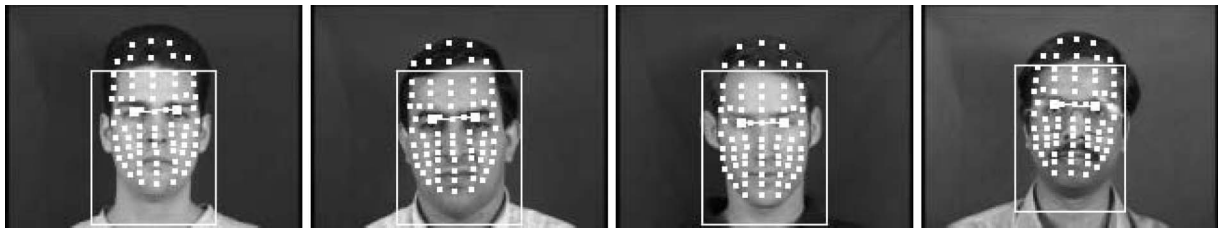

Figure 18. Facial feature extractions using XM2VTS database.

Figure 19. Mouth feature extractions using XM2VTS database.

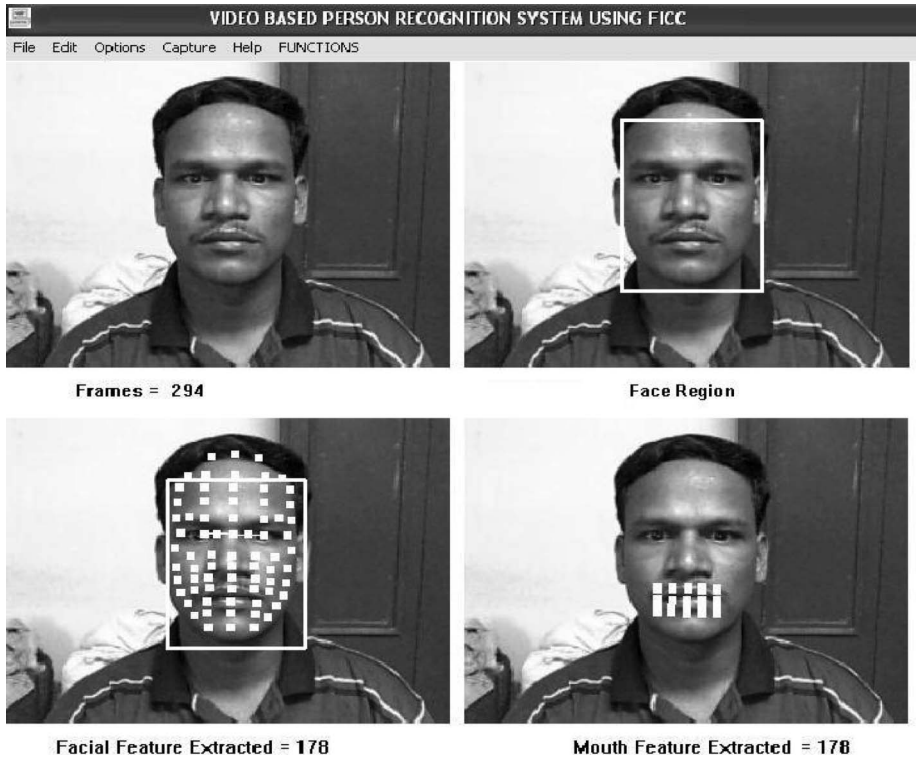

Figure 20. Snapshot of the the video-based person recognition system using FICC.

processing unit. The lighting conditions in the laboratory are not controlled, and hence there is a slight increase in the EER. The performance of the recognition system must be invariant to the size of the mouth, background, orientation and pose of the face and lighting conditions, in order to use it for commercial applications. The proposed method can be extended and used for applications such as Internet access and secured automated teller machine (ATM) transactions in controlled environment. The proposed method is not sensitive to the size of the mouth, its position in the image and its background, and orientation of the face.

It is also not sensitive to the pose of the face as long as the eye regions are visible. The proposed method is less sensitive to variation in the image brightness. However, the proposed method is sensitive to shadows, variation in lighting conditions and profile view of the face. Person recognition in uncontrolled lighting conditions and real time applications is still a challenging task.

7. Conclusion

In this article, a method was proposed for automatic person recognition using FICC. The proposed method uses the method proposed by Viola and Jones (2001) to detect the face region, and the face region is processed in $Y\ C_bC_r$ colour space to determine the locations of the eyes. The centre of the mouth is determined relative to the locations of the eyes. The facial features are extracted relative to the locations of the eyes, and mouth features are extracted relative to the locations of the eyes and mouth using multiscale morphological operations. FICC and exclusive-OR operation for matching are used to recognise a person in video sequences using face and mouth modalities. The performance of the system using FICC is evaluated in real time in the laboratory environment, and the system achieves an RR of 97.0% and an EER of about 0.89% for 50 subjects. The performance of the system is also evaluated for XM2VTS database, and the system achieves an RR of 100% and an EER of about 0.27%. PFICC and Euclidean distance for matching are also used to recognise a person in video sequences using face and mouth modalities. The performance of the system using PFICC is evaluated in real time in the laboratory environment, and the system achieves an RR of 99.0% and an equal error rate (EER) of about 0.84% for 50 subjects. The performance of the system using PFICC is also evaluated for XM2VTS database, and the system achieves an RR of 100% an EER of about 0.23%. The method is invariant to the size of the mouth, its position in the image and its background. The feature extraction techniques are computationally efficient, and the system recognises a subject within a reasonable time. Person recognition in uncontrolled lighting conditions and real time applications is still a challenging task. The proposed method can be extended and used for applications such as Internet access and secured ATM transactions in controlled environment.

References

Bartlett, M.S., Lades, H.M., and Sejnowski, T.J., 1998. Independent component representations for face recognition. *Proceedings of the SPIE, Conference on human vision and electronic imaging III*, Vol. 3299. San Jose, CA, 528–539.

Belhumeur, P., Hespanha, J., and Kriegman, D., 1997. Eigenfaces *vs*. fisherfaces: recognition using class specific linear projection. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (7), 711–720.

Chaudhari, U., *et al*., 2003. Audio-visual speaker recognition using time-varying stream reliability prediction. *IEEE international conference on acoustics, speech and signal processing (ICASSP '03)*, Vol. 5, Hong Kong, 712–715.

Chen, T., 2001. Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18 (1), 9–21.

Fleuret, F. and Geman, D., 2001. Coarse-to-fine face detection. *International Journal of Computer Vision*, 41 (1–2), 85–107.

Gao, Y. and Leung, M., 2002. Face recognition using line edge map. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (6), 764–779.

Gross, R. and Shi, J., 2001. The CMU motion of body (MoBo) database. Technical Report, CMU-RI-TR-01-18. Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Heisele, B., Alessandro, V., and Poggio, T., 2002. Learning and vision machines. *Proceedings of the IEEE*, 90 (7), 1164–1177.

Howell, A.J. and Buxton, H., 1996. Towards unconstrained face recognition from image sequences. *In*: *Proceedings of the international conference on automatic face and gesture recognition (FG '96)*, Killington, Vermont, 224–229.

Hsu, R., Abdel-Mottaleb, M., and Jain, A., 2002. Face detection in color images. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (5), 696–706.

Jackway, P. and Deriche, M., 1996. Scale-space properties of the multiscale morphological dilation-erosion. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18 (1), 38–51.

Jourlin, P., *et al*., 1997. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18, 853–858.

Kakadiaris, I.A., *et al*., 2007. Three-dimensional face recognition in the presence of facial expressions: an annotated deformable model approach. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (4), 640–648.

Kanak, A., *et al*., 2003. Joint audio–video processing for biometric speaker identification. *IEEE international conference on acoustics, speech and signal processing (ICASSP '03)*, Hong Kong, 377–380.

Kotropoulos, C. and Pitas, I., 2001. Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (7), 735–746.

Kotropoulos, C., Tefas, A., and Pitas, I., 2000. Frontal face authentication using discriminating grids with morphological feature vectors. *IEEE Transactions on Multimedia*, 2 (1), 14–26.

Lades, M., *et al*., March 1993. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 43 (3), 300–311.

Lam, K. and Yan, H., 1996. Locating and extracting the eye in human face images. *Pattern Recognition*, 29 (5), 771–779.

Leung, S., Wang, S., and Lau, W., 2004. Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Transactions on Image Processing*, 13 (1), 51–62.

Liu, X. and Chen, T., 2003. Video-based face recognition using adaptive hidden Markov models. *In*: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2003)*, Vol. 1, Madison, Wisconsin, 340–345.

Messer, K., *et al*., 1999. XM2VTSDB: extended M2VTS database. *Proceedings of the second international conference on audio and video based biometric person authentication (AVBPA 99)*, Washington, DC, 72–77.

Mian, A.S., Bennamoun, M., and Owens, R., 2007. An efficient multimodal 2D–3D hybrid approach to automatic face recognition. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (11), 1927–1943.

Nikolaidis, A. and Pitas, I., 2000. Facial feature extraction and pose determination. *Pattern Recognition*, 33 (11), 1783–1791.

O'Toole, A.J., *et al*., 2007. Face recognition algorithms surpass humans matching faces over changes in illumination. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (9), 1642–1646.

Rowley, H., Baluj, S., and Kanade, T., 1998. Neural network-based face detection. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (1), 23–38.

Savvides, M., Kumar, B.V., and Khosla, P., 2002. Face verification using correlation filters. *Proceedings of the 3rd IEEE automatic identification advanced technologies*, Tarrytown, NY, 56–61.

Sim, T., Baker, S., and Bsat, M., 2002. The CMU pose, illumination, and expression (PIE) database. *In*: *Fifth IEEE international conference on automatic face and gesture recognition*, 20–21 May 2002. Washington, DC: IEEE Computer Society Digital Library, 46–51.

Smeralsi, F., Carmona, O., and Bigun, J., 2000. Saccadic search with Gabor features applied to eye detection and real-time head tracking. *Image and Vision Computing*, 18 (4), 323–329.

Steffens, J., Elagin, E., and Neven, H., 1998. Personspotter-fast and robust system for human detection, tracking and recognition. *In*: *Proceedings of the international conference on automatic face and gesture recognition*, Nara, Japan. Washington, DC: IEEE Computer Society, 516–521.

Swets, D. and Weng, J., 1996. Using discriminant Eigen features for image retrieval. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18 (8), 831–836.

Terrence, C., *et al*., 2006. Total variation models for variable lighting face recognition. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (9), 1519–1524.

Timo, A., Abdenour, H., and Matti, P., 2006. Face description with local binary patterns: application to face recognition. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (12), 2037–2041.

Turk, M. and Pentland, A., 1991. Eigenfaces for recognition. *Journal of Cognitive NeuroScience*, 3, 71–86.

Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. *In*: *IEEE international conference on computer vision and pattern recognition (CVPR 2001)*, Kauai Marriott, Hawaii, Vol. 1, 511–518.

Wang, S., *et al*., 2004. Lip segmentation with the presence of beards. *IEEE international conference on acoustics, speech and signal processing (ICASSP '04)*, Vol. 3, Montreal, Canada, 529–532.

Wiskott, L., Fellous, J., and Malsburg, C., 1997. Face recognition by elastic bunch graph matching. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (7), 775–779.

Xiang, C., Fan, X.A., and Lee, T.H., 2006. Face recognition using recursive fisher linear discriminant. *IEEE Transactions on Image Processing*, 15 (8), 2097–2105.

Yang, M., Kriegman, D., and Ahuja, N., 2002. Detecting faces in images: a survey. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (1), 34–58.

Zhao, W., *et al*., 2003. Face recognition: a literature survey. *ACM Computing Surveys*, 35, 399–459.

Zhao, W., *et al*., 2000. Face recognition: a literature survey. *ACM Computing Surveys*, 35 (4), 399–459.

Zhou, S., Krueger, V., and Chellappa, R., 2003. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91, 214–245.