Face Recognition in Video Based on One Frontal View

Marijan Zivic
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
E-mail: marijan.zivic@gisdata.com

Abstract — Many face recognition techniques use image sequence as input for pose invariant face recognizer. But what if there is only one frontal image of person like the ones defined by biometric norm that can be seen on passports and other documents. This paper presents a new technique in face recognition in video (FRiV) using a single image per person as training data. Because of one example view technique was tested under same controlled lighting and with the same camera. The novelty of this paper is using Elliptical and Circular Gabor Filter to find accurate position of persons' left and right eye on faces in video clips. Accurate eye location is essential parameter in creating normalized face images from video clip frames suitable for recognition process. In this paper recognition is realized using eigenfaces. Recognition results on video clip depend on how accurate the frames with face are normalized.

1. INTRODUCTION

Over the last twenty years automatic face recognition from still images and video clips has been active search area. Robust solution of face recognition could be used in a wide range of potential applications such as biometric identity verification. Even though numerous face recognition algorithms have been proposed, robust system that can be used in real world situations and applications has not yet been developed because such objective is very challenging. The challenges come from large variations in visual information due to illumination conditions, head pose, facial expressions, aging and disguises and occlusions such as facial hair, glasses, or cosmetics.

Application of face recognition as biometric identity verification was inspiration for building system described in this paper. Successful development of such system could make verification in security systems more efficient and reliable. The main characteristic of such system is that in most cases it has only one image for each person that can be used for training. Dimensions and proportions of input face images are defined by biometric norm [1]. Recently, there has been some work on face recognition in video by using video sequences as training data [3]. Since the system described in this paper considers only one face image per person, recognition is achieved through standard procedure of Principal Component Analysis (PCA), or K-L Transform [2]. Video clips used for testing are taken under the same lighting condition as training images. Training set of face images was normalized using manually marked eye locations. Faces in video frames should be normalized as close as possible to training data. Accurate normalization considerably improves recognition results on video frames containing faces with frontal pose. In order to normalize video frames automatic eye location detection is needed.

Eye and iris center automatic detection method is based on Ring Gabor Filter [4]. First the eye region is got by an Elliptical Gabor Filter (ELF). Then the centre of eyeballs is found by the Circular Gabor Filter (CGF).

Only the frames with both irises detected are used for recognition. That eliminates some poses that would not be successfully recognized. Assumption is that the difference between video frames with full frontal and non frontal pose will be big enough not to influence recognition efficacy. The goal is to achieve that the similarity measures of frontal pose frames are considerably larger than the similarity measures of frames that do not represent frontal pose. In that case only frames containing frontal poses will be considered in determining identity of person on video.

2. FACE IMAGE NORMALIZATION

The goal of face image normalization is to transform image into standard format that removes variations that can affect recognition performance. Normalization consists of two steps: geometric normalization and masking. The goal of geometric normalization is to remove variations in size, orientation and location of face in the image. Masking step hides background of the face that can contain features like clothes, hair and background that affect recognition. Image normalization is used in two parts of the system: for creating training images for PCA recognition method and for transforming video frames into probe images for recognition process.

In order to perform normalization, locations of both eyes must be defined. Normalization process goes as follows:

1. Rotate image so that eyes are aligned.
2. Scale image so that the distance between eyes is set to 40 pixels.
3. Clip image to size 110×110, so that the right eye position is (35, 33) and left eye is (75, 33).
4. Mask image with ellipsoid mask. Center of ellipsoid is the middle of the image. Width of ellipsoid is 60% of image width and height of ellipsoid is 110% of height of image.

Illumination normalization is not used because this paper describes recognition system under controlled light and camera. Figure 1 shows normalization for training and probe images.
3. THE EIGENFACES METHOD

Training images are represented as two-dimensional N by N array of (8-bit) intensity values. Images can also be considered as very long vectors of dimension N^2. Representation of images as vectors represents a new huge space called image space where each image represents a point. Since faces have similar structure they will not be randomly distributed in image space and can be described by a relatively low dimensional subspace.

The main idea of PCA is to find the set of directions in the images space that best account for the distribution of face images. In other words, to find vectors in image space along which the variance of the cluster is maximum. These new vectors define the subspace of the face images called face space. Transformation of vectors of images to face space we express images as linear combination of base images. These base images have same dimensionality as training images and because they look like faces are called eigenfaces. Recognizing similar faces is equivalent to calculating the nearest point to the novel face in face space.

1.1. Calculating Eigenfaces

Direction of variance in a high dimensional space can be extracted from the covariance matrix of the data points. The eigenvectors of the covariance matrix point in directions of maximum variance of the data. Mean square error between original and image transformed to face space is minimized by selecting the eigenvectors with largest eigenvalue.

Let the training set of face images be Γ₁, Γ₂, Γ₃,....,Γₖ. The average face for this set of images is defined by:

\[ \Psi = \frac{1}{M} \sum_i \Gamma_i, i = 1, ..., M \]

(1)

Each face differs from the average by:

\[ \Phi_i = \Gamma_i - \Psi \]

(2)

The covariance matrix of the data is defined as:

\[ C = \frac{1}{M} \sum_i \Phi_i \Phi_i^T = AA^T, i = 1, ..., M \]

(3)

where \( A = [\Phi_1, \Phi_2, ..., \Phi_M] \) and \( C \) has dimension \( w*h \times w*h \) where \( w \) is the width of the image and \( h \) is the height. Calculating \( w*h \) eigenvectors and eigenvalues for matrix \( C \) is computationally demanding. Computationally feasible method for finding eigenvectors and eigenvalues is needed. If the number of training images is less than the dimension of image space (\( M < N^2 \)) there will be only \( M-1 \) eigenvectors enough for recognition process. If we consider \( V_i \) to be the eigenvectors of matrix \( A^T A \) where \( A^T A \) is only \( M \times M \) matrix, then:

\[ (A^T A)V_i = \lambda_i V_i \]

(4)

where \( \lambda_i \) are the eigenvectors, then

\[ A(A^T A)V_i = A(\lambda_i V_i) \]

(5)

which means

\[ (AA^T)(AV_i) = \lambda_i (AV_i) \]

(6)

and \( AV_i \) are the eigenvectors of \( C = AA^T \). Therefore, the eigenvalues of \( C \) are given by:

\[ U_i = AV_i = \sum_i v^i_k \Phi_k, k = 1, ..., M \]

(7)

Where \( (k = 1, ..., M - 1) \) and \( v^i_k \) is the \( k \)th element of \( V_i \). All of the images in training set are projected onto the eigenvectors by:

\[ P_k = U_k^T (\Gamma - \Psi), k = 1, ..., M' \]

(8)

Where \( \Psi \) is the average face image given by equation (1), \( M' < M-1 \). For each image in the training set a pattern vector defined as \( P_k = [p_1, ..., p_M] \) is calculated. Pattern vector describes the contribution of each eigenface in representing the input face image. Eigenvectors and pattern vectors of each image of the training set are then stored so that they don’t have to be calculated in each recognition process.

1.2. Using Eigenfaces to Classify a Face Image

Let Γ be a new face image. Image is projected onto the eigenvectors using equation 8) to get pattern vector \( P = [p_1, ..., p_M] \). The simplest method for determining which face class provides the best description of an input face image is to find the face class that minimizes the Euclidian distance:

\[ \epsilon_k = \| P - P_k \|^2 \]

(9)

where \( P_k \) are pattern vectors for training images.
4. EYE LOCATION

In this paper Ring Gabor Filters (RGF) are used for determining eye location. The shape of eye is nearly elliptical and the shape of eyeball is circular. First, the eye region is detected using EGF. Then the center of eyeball is detected using CGF. Traditional Gabor filters have been used for texture segmentation and classification, fingerprint recognition and face recognition. RGF is the modification of Traditional Gabor Filter (TGF). Ring filters were originally designed and used by Coggins and Jain [5] for texture analysis early in the 1980s. RGF, including EGF and CGF has less parameters and simpler structure than traditional Gabor filter. It is more matchable with the eye shape.

4.1. Elliptical Gabor Filter (EGF)

The shape of EGF matches ellipse. The form of EGF is similar to traditional Gabor filter, both of which are Gaussian functions modulated by the sine function. The difference is in shape of the sine function. EGF is defined as:

\[
G(x, y) = g(x', y')e^{2\pi F \sqrt{x'^2 + y'^2}}
\]  

where

\[
(x', y') = (x \cos \theta + y \sin \theta, -x \sin \theta + y \cos \theta)
\]  

F is the spatial central frequency of the filter in the frequency domain, \(g(x, y)\) is 2D Gaussian envelope given by:

\[
g(x, y) = \frac{1}{2\pi \sigma_x \sigma_y} e^{-\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right)}
\]

where \(\sigma_x\) and \(\sigma_y\) are variances (scale factors) along x and y axes respectively. \((x, y)\) is the center of the receptive field in the spatial domain, \(\theta\) is the rotation angle of the filter.

The horizontal EGF can stand out the eye region efficiently wherever the image is frontal or in slight rotating. So when designing the EGF, only three parameters \(\sigma_x\), \(\sigma_y\) and F are considered. This simplifies the filter to some degree. Values for EGF parameters used in this paper are:

\[
\sigma_x = 3.7, \sigma_y = 8, \quad F = 0.3
\]

Figure 2 shows process of locating eye regions on faces video frames detected by using Haar cascades introduced by Viola [7] and implemented in OpenCV [6]. Locating eye regions goes as follows:

1. Filter the face region detected by Haar cascades implemented in OpenCV using EGF.
2. Binarize the filter output using threshold.
3. Apply morphological dilation.
4. Calculate the centroid of white region.
5. If there are more then one white region choose one that is closest to the middle of the picture.

5. If there are more than one white region choose one that is closest to the middle of the picture.

4.2. Circular Gabor Filter (CGF)

Circular Gabor filter removes orientation selectivity of TGF. Orientation selectivity could represent disadvantage in case of detecting eyeball which is orientation invariant shape. The CGF is defined as:

\[
G(x, y) = g(x', y')e^{2\pi F \sqrt{x'^2 + y'^2}}
\]  

Where \(F\) is the radial frequency and \(g(x, y)\) is 2-D Gaussian envelope assumed to be isotropic as follows:

\[
g(x, y) = \frac{1}{2\pi \sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}}
\]

When designing CGF only two parameters are considered: \(\sigma\) and \(F\). Values for CGF parameters used in this paper are:

\[
\sigma = 3.7, \quad F = 0.1445321
\]

Figure 2 shows process of locating eyeballs on eye region images that are extracted using EGF. Locating eye regions goes as follows:

1. Filter the eye region using CGF filter.
2. Binarize the filter output using threshold.
3. Apply morphological dilation.
4. Calculate the centroid of white region.
5. If there are more than one white region choose one that is closest to the middle of the picture.

5. Calculate the centroids of white regions.
6. Segment the eye regions to get two rectangles for further eyeball detection.
5. EXPERIMENTS

Images and video clips from 23 persons are used for experimenting. Images were taken with the same camera and lightning conditions as video clips. Images were used for eigenfaces training. Training images were normalized as described in section 2. Video clips captured persons while talking and had full freedom of head movement as long as their body was in vertical position. Distance of persons from camera was about 1 meter. Resolution of video clips is 720x576. Video clips consist of frames with different head poses and face expressions. Each person was marked with ID=1,...,23.

Threshold value for distances in recognition process has been experimentally determined. Because of different head pose and expressions in video clips there will be considerable number of frames that will not be recognized or will be recognized as persons that do not represent true identity. Frame distances are grouped by the persons that the frames are associated to. For each video-clip, the following statistics were computed:

1. F1: The number of frames in a video clip, in which face is unambiguously recognized.
2. F2: The number of frames in a video clip which a are associated with someone else. This is a worst case result but it does not have to be bad. In most cases this kind of result is the outcome of the recognition of frame containing face in a non frontal pose. It could also be outcome of recognition of face that was not properly normalized due to the eyeball not detected accurate enough. Under some pose and light conditions RGF does not detect eyeball accurately.
3. F3: The number of frames in which the face is not associated with any of the training faces. This result happens if the distance of face in frame from training images in eigenspace is greater than the given threshold.
4. D11: Average distance of frames that were successfully recognized.
5. D12: Minimal distance of frames that were successfully recognized.
6. D21: Minimal distance of average distances of groups of frames associated with the persons that do not represent true identity.
7. D22: Minimal distance of minimal distances of groups of frames associated with the persons that do not represent true identity.

Assumption is that the distance of full frontal faces captured from frames and accurately normalized will have the minimal distance to true identity in eigenspace. Most accurate voting mechanism should be found. Figure 4 shows that the number of successfully recognized frames, F1 can not represent similarity measure because it is often smaller than the number of frames not being successfully recognized, F2. Even if that number was greater, the system does not have the information about face pose so these results would be wrong in most cases. Frames that are not frontal will often give bad results. Figure 5 and Figure 6 show that both the average and minimal distances of frames that were associated with someone else are mostly greater than the distances of successfully recognized frames. This makes good base for proving the assumption described in the first chapter. Next experiment will compare voting for true identity based on average (AGD) and minimal (MGD) distances for group of frames associated with the same person.
TABLE I. RECOGNITION RESULTS

<table>
<thead>
<tr>
<th>Similarity measure</th>
<th>Number of clips</th>
<th>Recognition rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGD</td>
<td>23</td>
<td>52.2</td>
</tr>
<tr>
<td>MGD</td>
<td>23</td>
<td>73.9</td>
</tr>
</tbody>
</table>

Results are displayed in Table I. It can be seen that using minimal group’s distance give much better recognition rate.

It is evident from Figure 4 that some video clips have very low number of successfully recognized frames. There can be two reasons for that. First, video clips are not the same. Head positions, pose and face expressions are different on every clip. Some clips may appear without full frontal frames with satisfying quality. Second, in some conditions RGF does not detect eyes accurately and that disturbs the normalization process which is closely connected with recognition rate. The issues that can be noticed with RGF are:

- Detection of nose or eyebrow as eye region.
- Shadows around eyes reduce the detection of eyeballs.
- Glasses with thick frame in most cases will cause false eyeball detection.

5.1. Improving eyeball detection

Considering the described issues with eye location detection, slight improvements were done. To avoid detection of nose as eye region only upper part of face region will be examined. By analyzing face regions detected by Haar cascades method, a certain offsets in face region were experimentally determined to avoid interference with persons’ hair that could decrease threshold effect. Shadows around eyes and most of the glasses types are features that can be approximated with ellipsoid, so these features will also result as maximum after running EGF. To avoid that, CGF is applied without previous application of EGF. This works fine on upper face region clipped by determined offsets. Improvements have increased the number of frames with both eyes detected and detected eye locations were much closer to the real center of the eyeball. This results in better recognition performance.

Figure 7 shows improved eyeball detection procedure. Table II shows new recognition results of system with improved eyeball detection.

6. CONCLUSION

This paper presents face recognition system that recognizes faces from video clips based on one image per person defined by biometric norm. This kind of images can be found on passports and other document. The system was built for video records taken under controlled lightning and capturing device. Variance in faces caused by aging was not considered due to purchase frequency of documents with biometric images. Distances of video frames with non frontal face pose in eigenspace are greater enough from distances of frontal frames which makes good basis for recognition on test video set. This paper also presents application of RGF used for detection of eye location that is needed for video frame normalization. The overall recognition rate was improved with the improvement of the RGF eye location detection. The system should be combined with good pose detector in order to isolate only frames with faces in frontal pose or near enough to frontal pose.

REFERENCES