Heterogeneous Face Recognition and Synthesis Using Canonical Correlation Analysis

Mengyi Liu, Zhiyong Yuan, Yufeng Ma, Xingwei Chen, Qian Yin

1, First Author School of Computer, Wuhan University, Wuhan, 430072, P.R. China, liumengyi@whu.edu.cn
2, Corresponding Author School of Computer, Wuhan University, Wuhan, 430072, P.R. China, zhiyongyuan@whu.edu.cn
3, 4 School of Computer, Wuhan University, Wuhan, 430072, P.R. China, mayufeng@whu.edu.cn, cxweeiee@126.com
5 College of Information Science and Technology, Beijing Normal University, Beijing 100875, P.R. China

Abstract

Face images captured in different spectral bands are said to be heterogeneous. In this paper, we propose a new approach based on subspace-mapping for heterogeneous face recognition and synthesis. In the recognition section, Local Binary Pattern (LBP) is used as facial representation for near infrared (NIR), visual light (VIS) and 3D range images. Then Canonical Correlation Analysis (CCA) is applied to learn the mapping between the different LBP-face patterns. The corresponding matching scores are calculated in the CCA subspace for the final decision. In the synthesis section, according to the CCA transformation matrices obtained above, we apply ridge regression to determine an approximate linear relationship between the target pattern image and the projection vector of the probe. Experimental results are provided to evaluate the accuracy of the method. The work shows a practical solution for reliable heterogeneous face synthesis.

Keywords: heterogeneous face recognition, subspace-mapping, Canonical Correlation Analysis (CCA), Local Binary Pattern (LBP), linear regression

1. Introduction

Face recognition from images has been a hot research topic in computer vision because of its potential application values as well as theoretical challenges. The performance of the recognition can be affected by some external factors such as illumination, expression and posture. Reference [1] classified the images into VIS (visual light), NIR (near infrared), TIR (thermal infrared) according to the different spectral bands which the images are captured in. These different image types are said to be heterogeneous if they have different image formation characteristics. In the broad sense heterogeneous images can also be interpreted as the images coming from different sensors in VIS mode (like CCD camera, CMOS camera, etc.), or those under the same imaging model but different illumination or resolution. Face biometrics by matching between heterogeneous face images are collectively called heterogeneous face biometrics (HFBs).

For recent decades, face recognition in single pattern have been developed, especially for VIS and NIR. But the limitation is that both enrollment and query face images are assumed to be of the same type. However, in many important applications, face images can only be captured under visible light, such as video surveillance, photo-based identification and E-Passport, which are greatly influenced by illumination. NIR technology has been applied to accomplish the lighting invariant face recognition to some extent. By taking advantages of NIR images, and allow the use of existing ID face photos as gallery templates, the work will be more accuracy and efficiency.

There are two main approaches to heterogeneous face biometrics (HFB). The difference is whether to transform the problem to single pattern matching in the algorithm. Reference [2] proposed an analysis-by-synthesis method. In the framework a mapping model is constructed to transform face images from one type to another, and thereby perform the heterogeneous matching. However the
The synthetic process is pixel-based and time-consuming. Reference [3] and [4] analyze the local structure of normalized appearance, extract the common features of the two different types, and then the similarities between the extracted features are measured and compared with a certain threshold to decide the identify results. This kind of approach is efficiency for recognition but a little bit less intuitive than the former (because there is no synthesis-image as a “bridge” between two patterns). The method proposed in this paper is based on the second approach. The correlation projection is used to construct a common subspace for common “features” extraction. Combining an effective facial representation and the principle of partial similarity, the total performance is proved to be improved in our experiments.

The remaining part of the paper is organized as follows: In section 2, heterogeneous face recognition based on LBP and CCA is introduced. In section 3, we present a linear regression method for heterogeneous face synthesis. Experiments and analysis are conducted in section 4, followed by conclusion in the last section.

2. Heterogeneous face recognition

Although the heterogeneous face images from a same individual are significantly different in appearance, we can still achieve multi-modal patterns matching by image processing and transforming. In the entire process of heterogeneous face recognition, image representation and matching algorithm are two crucial parts.

2.1. Facial representation

Facial representation is the description and coding method of facial information. Generally it includes the presentation of geometric characteristics, algebra characteristics, fixed feature templates and eigenface, etc. An original facial representation is to transform the face matrix into a column vector. This method do not need complicate algorithm and remain the entire information of the image as well, however it always causes high dimension problem and overlooks the relationship between the neighboring pixels.

Local Binary Pattern (LBP), a non-parametric method, summarizes the local structures of an image efficiently [5]. The original LBP operator labels the pixels of an image by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary string, of which the corresponding decimal number is used for labeling. An illustration of the basic LBP operator is shown in Figure 1.

![Figure 1. An example of the original LBP operator](image)

Given a central pixel at \((x_c, y_c)\), the LBP results can be expressed in decimal form as:

\[
LBP(x_c, y_c) = \sum_{k=0}^{7} f(i_k - i_c) \cdot 2^k
\]  

(1)

Where \(k\) runs over the 8 neighbors of the central pixel, \(i_c\) and \(i_k\) are gray-level values of central pixel and surrounding pixels. The function \(f(x)\) is defined as:

\[
f(x) = \begin{cases} 
1 & \text{if } x \geq 0 \\
0 & \text{if } x < 0 
\end{cases}
\]  

(2)

According to Eqn. (1) and Eqn. (2), the LBP operator is invariant to the monotonic gray-scale transformations which preserve intensity order in local neighborhoods. The histogram of LBP labels calculated over a region is used as a texture descriptor.
To deal with the texture at different scales, the original LBP operator was extended to neighborhoods of different sizes. Local neighborhood is defined as a set of sampling points evenly spaced on a circle centered at the pixel to be labeled, and the sample points that do not fall in the pixels are expressed using bilinear interpolation, thus allowing any radius and number of sampling points in the neighborhood. Figure 2 shows some examples of the extended LBP operators; the notation \( (P, R) \) denotes a neighborhood of \( P \) sampling points on a circle of radius of \( R \).

![Figure 2. Examples of operators: circular (8, 1), (16, 2), and (8, 2)](image)

The LBP operator \( LBP_{(P, R)} \) produces \( 2^P \) different output values, corresponding to \( 2^P \) different binary patterns formed by the \( P \) pixels in the neighborhood [6]. The general idea for LBP is that a face image can be seen as a composition of micro-patterns which are described by the operator. But the histogram of LBP computed over the whole image encodes only occurrences of the micro-patterns without any indication about their locations. To also consider shape information of faces, the image can be divided into some local regions, from which LBP histograms are extracted and then concatenated into a single global one (containing both local texture and global shape information). Figure 3 shows some examples of the LBP-processed image and the corresponding histograms.

![Figure 3. Examples of LBP-processed faces (line 2) and the corresponding histograms (line 3), from left to right are VIS, NIR, 3D](image)

2.2. CCA based matching

Canonical correlation analysis (CCA) is a powerful tool for correlating two sets of multi-variant measurements in their leading factor subspaces [7]. Like Principal Components Analysis (PCA), CCA also reduces the dimensionality of the original signals, since only a few factor-pairs are normally needed to represent the relevant information; Unlike PCA, however, CCA takes into account the relationship between two signal spaces (in the correlation sense), which makes them better suited for regression task than PCA.

But in face recognition and other small sample size (SSS) problem, the traditional CCA has following disadvantages: 1) due to the singularity of the covariance matrices of its two groups of
features caused by the SSS problem, CCA fails if directly applied; 2) it cannot describe the nonlinear face recognition problem well for its globally linear property in nature. Thus we propose a patch-based CCA to eliminate the singularity problem and enhance the robustness to local variants.

Suppose the training data sets \( X = [x_1, x_2, \ldots, x_N] \), \( Y = [y_1, y_2, \ldots, y_N] \) (or \( \{(x_i, y_i), \ldots, (x_N, y_N)\} \)) are \( N \) pairs of heterogeneous faces of \( C \) people. For each sample \( x_i \) (or \( y_i \)), the face image is divided into \( m \) overlapped patches as \( x_i = [x'_1, x'_2, \ldots, x'_m] \) (or \( y_i = [y'_1, y'_2, \ldots, y'_m] \)). The corresponding patches (in the same location of the heterogeneous images) of the \( N \) samples constitute \( m \) sub-patterns as \( \{X'_1, X'_2, \ldots, X'_N\} \) (or \( \{Y'_1, Y'_2, \ldots, Y'_N\} \)), then we obtain the patch-based CCA training pairs:

\[
\begin{align*}
X' &= [x'_1, x'_2, \ldots, x'_m] \\
Y' &= [y'_1, y'_2, \ldots, y'_m]
\end{align*}
\]

Figure 4 is an example of dividing a face image into \( m \) overlapped patches, where \( m = 25 \) and \( overlap = 0.25 \). Figure 5 shows the way of constructing the patch-based CCA training pairs.

The leading factor subspaces are the linear subspaces of the training data sets \( X \) and \( Y \), both of a reduced dimensionality \( d \). CCA takes into account the two data sets simultaneously and finds the optimal linear projective matrices, also called canonical projection pairs:

\[
\begin{align*}
W'_X &= [w'_1, w'_2, \ldots, w'_d] \\
W'_Y &= [w'_1, w'_2, \ldots, w'_d]
\end{align*}
\]

Such that \( x'_i = w'_x^T \cdot X' \) and \( y'_i = w'_y^T \cdot Y' \) are most correlated. This is done by maximizing the following correlation:
\[ \rho(w_x^j, w_y^j) = \frac{E[x_x^j y_y^j]}{\sqrt{E[\|x_x^j\|^2]E[\|y_y^j\|^2]}} = \frac{w_x^j C_y w_y^j}{\sqrt{w_x^j C_y w_y^j C_x w_y^j}} \]  

(5)

Subject to

\[
\begin{align*}
&w_x^j C_y w_y^j = w_y^j C_x w_x^j = 1 \\
&w_x^j C_y w_y^j = w_y^j C_x w_x^j = 0 \\
&s, t = 1, 2, ..., d, \ s \neq t \\
&w_x^j \in R^d, w_y^j \in R^d
\end{align*}
\]

(6)

Where \( C_x \), \( C_y \) and \( C_{xy} \) are the correlation matrices computed from the training data sets \( X' \) and \( Y' \). According to CCA \( w_x^j \) and \( w_y^j \) satisfy

\[
\begin{align*}
& C_x^{C_y} C_y^{C_x} W_x^j = C_x^{C_y} W_y^j, \Lambda^j \\
& C_y^{C_x} C_x^{C_y} W_y^j = C_y^{C_x} W_x^j, \Lambda^j
\end{align*}
\]

(7)

The problem can be converted to the following generalized eigenproblem:

\[ A w = \lambda B w \]

(8)

Where

\[
A = \begin{pmatrix} C_x & 0 \\ C_y & 0 \end{pmatrix}, \ B = \begin{pmatrix} C_x & 0 \\ 0 & C_y \end{pmatrix}, \ w = \begin{pmatrix} W_x^j \\ W_y^j \end{pmatrix}
\]

(9)

The solution \( W_x^j \) and \( W_y^j \) can be found using singular value decomposition [8]. Thus we obtain \( m \) sub-pattern canonical projection matrices pairs for \( m \) overlapped patches.

The next step evaluates correlation between the testing data sets \( X \) and \( Y \) in the learned subspaces. The testing sets are also transformed into \( m \) sub-patterns using the same method in training sets. Features \( X_{\alpha} \) and \( Y_{\alpha} \) are extracted from the patch \( k \) and they are projected into the CCA subspace as \( x_{\alpha}^k \) and \( y_{\alpha}^k \) using \( W_x^j \) and \( W_y^j \) found above. The correlation between \( x_{\alpha}^k \) and \( y_{\alpha}^k \) is calculated as the matching score of patch \( k \).

\[
corr(x_{\alpha}^k, y_{\alpha}^k) = x_{\alpha}^k y_{\alpha}^k / \| x_{\alpha}^k \| \| y_{\alpha}^k \|
\]

(10)

For each testing pair, the \( m \) correlation values of \( m \) patches are sorted in descending order. The top \( \alpha \cdot m \) values are selected to calculate the arithmetic mean as the final similarity of the whole face (where \( \alpha \) is a proportion factor ranged from 0 to 1 and \( \alpha \cdot m \) must be an integer). Finally the similarities are measured and compared with a certain threshold to decide the identify results. Figure 6 shows the process of the CCA based matching between VIS and NIR faces.
3. Heterogeneous face synthesis

After employing CCA on two data sets, we can extract the most correlative component pairs from the original data. Denote the patch based sample pairs from the data sets by random vector $x$ and $y$. Let $x' = W_x x$, where $W_x$ is the CCA transformation matrix, so $x'$ is the most correlative components of $x$ to $y$. Our purpose is to learn the mapping between $x'$ and $y$ [9].

Introduce a regression parameter matrix $R$ to describe the relationship. Specifically, we assume $y$ and $x'$ have a linear relationship as:

$$y = R x' + \delta$$

(11)

Where $\delta$ is the noise item which obeys the Gaussian distribution, $\delta \sim N(0, \sigma^2 I)$, where $I$ is the identity matrix. Thus we have:

$$sim(y \mid x', R) = \frac{1}{\text{norm}} \exp\left\{-\frac{(y - R x')^\top (y - R x')}{2\sigma^2}\right\}$$

(12)

Where $\text{norm}$ is the normalization coefficient. By maximizing the function in the training set with respect to $R$

$$R' = \arg \max_x \left\{-\frac{1}{2\sigma^2} \sum_i (y_i - R x'_i)^\top (y_i - R x'_i)\right\}$$

(13)

$$= \arg \min_x tr((Y - RX')(Y - RX')^\top)$$

Where $i$ is the index of patches ranging from 1 to $m$. And we can get the solution by setting the derivative of objective function (with regard to $R$) to zero as

$$R^* = XX' (X' X')^{-1}$$

(14)

Moreover, in order to improve the generalization of result, we can impose adjustment factor onto the log-likelihood function $sim(y \mid x', R)$ according to prior knowledge.
Where $\lambda$ controls the trade-off between the accuracy in the training set and the generalization. We can then obtain the optimal result as

$$R^* = \arg \max_{R} \text{tr}((Y - RX)^{(Y - RX)^T + \lambda RR^T})$$

(15)

(16)

Which is essentially equivalent to the ridge regression [10].

Given a testing sample $x_{test}$, $x_{test}'$ is the computed using CCA transformation matrix

$$x_{test}' = w^T x_{test}$$

(17)

And the synthesis $y_{syn}$ is then obtained by

$$y_{syn} = R' x_{test}'$$

(18)

The corresponding diagram of the synthesis is given in Figure 7.

\[\begin{align*}
R' &= \arg \max_{R} \text{tr}((Y - RX)^{(Y - RX)^T + \lambda RR^T}) \\
\text{Where } \lambda &\text{ controls the trade-off between the accuracy in the training set and the generalization. We can then obtain the optimal result as } R' = YX^T (X^T X^T + \lambda I)^{-1} \\
\text{Which is essentially equivalent to the ridge regression [10].}
\end{align*}\]

Given a testing sample $x_{test}$, $x_{test}'$ is the computed using CCA transformation matrix

$$x_{test}' = w^T x_{test}$$

(17)

And the synthesis $y_{syn}$ is then obtained by

$$y_{syn} = R' x_{test}'$$

(18)

The corresponding diagram of the synthesis is given in Figure 7.

\[\begin{align*}
\text{Figure 7. Synthesis of a NIR face image from VIS face image}
\end{align*}\]

According to the synthesis result, we can further improve the recognition performance. Let $y^* = y_{syn}$ and $y = y_{test}$. Define the matching score as

$$\text{score}(y^*, y) = \frac{1}{\sqrt{(y - y^*)' (y - y^*) / d}}$$

(19)

Where $d$ is the dimension of the vector $y$. Finally we can fusion the matching score of each patch in the same way as in recognition phase.

4. Experiments

In our experiments, all the images are from the HFB Face Database of Chinese Academy of Science. The NIR and VIS face image pairs of 100 persons was captured in different time, containing variations in pose, expression and illumination. For each person, there are 4 NIR and 4 VIS mutually registered. All the images are normalized into the size of 128×128 pixels.

We choose 200 pairs of samples as the training space, including 50 persons with 4 images of each. Correspond to different facial representation, different matching method, we have several groups of comparative trials as following.
(a) We evaluate two ways of image representation, one is just transforming the face matrix into a column vector (which is said to be “direct representation”), the other is LBP-based presentation which reduces the feature dimension to 256.

(b) Considering the size of original face image, we evaluate three scales of the patch. The corresponding parameters are shown in Table 1.

Table 1. The parameters in different scales of the patch

<table>
<thead>
<tr>
<th>Patch size</th>
<th>Overlap</th>
<th>Patch num</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>20x20</td>
<td>10%</td>
<td>7x7(49)</td>
<td>400/256(LBP)</td>
</tr>
<tr>
<td>32x32</td>
<td>25%</td>
<td>5x5(25)</td>
<td>1024/256(LBP)</td>
</tr>
<tr>
<td>128x128</td>
<td>0%</td>
<td>1x1(1)</td>
<td>16384/256(LBP)</td>
</tr>
</tbody>
</table>

(Comment: Suppose the patch size is L*L, overlap is d%, patch num is N*N. There is L*(1-d%)*N+L*d%=128.)

The third line represents the situation that extracting the global features from each whole image without division. Method using “direct representation” on this situation is not evaluated due to the memory limitation.

(c) The matching method using linear regression and without linear regression are also compared in the experiments.

Table 2 provides verification rates at FPR=0.1, FPR=0.01 and FPR=0.001. The ROC (Receiver Operator Characteristic) curves for 8 algorithms are shown in Figure 8 (where LR represents linear regression). We use the order in the legend.

Table 2. Comparison of verification rates for the 8 algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPR=0.1</td>
<td>64.69%</td>
<td>85.19%</td>
<td>98.88%</td>
<td>91.81%</td>
<td>17.31%</td>
<td>93.06%</td>
<td>93.69%</td>
<td>49.38%</td>
</tr>
<tr>
<td>FPR=0.01</td>
<td>47.44%</td>
<td>61.31%</td>
<td>92.31%</td>
<td>69.63%</td>
<td>2.38%</td>
<td>75.13%</td>
<td>82.31%</td>
<td>23.96%</td>
</tr>
<tr>
<td>FPR=0.001</td>
<td>41.19%</td>
<td>42.00%</td>
<td>81.38%</td>
<td>49.19%</td>
<td>0.31%</td>
<td>57.81%</td>
<td>66.06%</td>
<td>12.06%</td>
</tr>
</tbody>
</table>

From Table 2, we can obtain following inferences.

(a) No matter what face representation algorithm is used, the global scale achieved poor recognition rates.

(b) Both CCA+LBP and CCA+LBP+LR achieved higher rates than single CCA.

(c) When patchsize=20, CCA+LBP is better than CCA+LBP+LR. When patchsize=32, CCA+LBP+LR is better than CCA+LBP. That means whether linear regression is useful depends on the actual experimental conditions (e.g. image resolution, patch numbers, representation algorithm, etc.)
5. Conclusion

In the paper, we introduce the characteristic of the HFB technology and the research significance. By analyzing the current solutions, we proposed a novel recognition algorithm based on
subspace-mapping for heterogeneous face recognition and synthesis. The method takes the CCA subspace as a medium to match the heterogeneous images in database, and further improve the recognition accuracy. In addition, introducing the LBP representation and patch based method emphasize the local structure features and eliminate the high-dimension problem as well. However the algorithm depends on the linear mapping between the images, the preprocessing is crucial and the pixel alignment will significantly affect the experimental results. Future work will focus on this problem to make the recognition algorithm more robust and efficient.

6. Acknowledgements

This work was supported by the National Basic Research Program of China (Grant no. 2011CB707904) and the Project of Scientific Research & Innovative Experiment of National Undergraduate (Grant no. 101048624).

7. References