Person-Specific Face Detection in a Scene with Optimum Composite Filtering and Colour-Shape Information

Seokwon Yeom1,* and Yong-Hyun Woo1

1 Division of Computer and Communication Engineering, Daegu University, Gyeongsan, South Korea
* Corresponding author E-mail: yeom@daegu.ac.kr

Received 20 Jun 2012; Accepted 10 Oct 2012

© 2013 Yeom and Woo, licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract Face detection and recognition have wide applications in robot vision and intelligent surveillance. However, face identification at a distance is very challenging because long-distance images are often degraded by low resolution, blurring and noise. This paper introduces a person-specific face detection method that uses a nonlinear optimum composite filter and subsequent verification stages. The filter’s optimum criterion minimizes the sum of the output energy generated by the input noise and the input image. The composite filter is trained with several training images under long-distance modelling. The candidate facial regions are provided by the filter’s outputs of the input scene. False alarms are eliminated by subsequent testing stages, which comprise skin colour and edge mask filtering tests. In the experiments, images captured by a webcam and a CCTV camera are processed to show the effectiveness of the person-specific face detection system at a long distance.

Keywords Face identification, face detection, correlation filtering, composite filtering, skin colour test, edge mask

1. Introduction

Recently, the use of CCTV systems has dramatically increased for various military and civilian applications, thus security can be improved effectively by searching the identity of the faces captured on the CCTV images. However, intelligent surveillance, such as face identification at a distance, is very challenging and still preformed manually, which requires much time and labour. The most common difficulties with face recognition are unexpected changes in pose, illumination, and expression [1]. Further, images acquired at a distance are often degraded by low resolution, out-of-focusing and motion blurs, and noise [2,3].

A popular face detection method is the AdaBoost algorithm [4,5], however, it often requires a large computational time for the searching process. For face recognition, statistical pattern recognition methods, such as Fisher discriminant analysis and principal component analysis, are commonly used [6]. Gabor filtering with feature extraction is a promising alternative [1], but this method generally requires high resolution images.
Correlation filtering for face identification has been researched in [7-9].

In this paper, we discuss person-specific face detection in the scene that uses a nonlinear composite filter with the long distance observation model and subsequent verification stages [10]. The nonlinear composite filter is a correlation approach which can identify a face of a specific person in the scene. Since this method does not require extraction and alignment of facial regions, a compact system can be constructed to determine the presence of a specific person and his/her location in the scene [7-9]. Further, the computational load of the correlation filter is relatively low since the filtering can be processed in the frequency domain. The correlation filters have been developed with shift-invariant, output-normalized and noise-robust properties [8,9].

The optimum criterion of the filter minimizes the sum of the output energy functions generated by the input scene noise, for noise robustness, and by the input image, for high discriminant capability [11]. This composite filter has been developed for automatic target recognition (ATR) dealing with background and overlapping noise in the scene [11]. In this paper, the performance of the composite filter is improved by applying the long-distance observation model to a high quality training image set.

One facial region may be selected with the highest peak of the filtered outputs [12], however, considering inevitable non-zero false alarm rates, multiple outputs are embraced as candidate facial regions. Those candidate windows are verified through the subsequent testing stages, which comprise the skin colour and edge mask filtering tests; first, the regions output by the filter are subject to the skin colour test and then an edge-mask is applied to the edge image of the candidate window.

In the experiment and simulation, our algorithm is tested using a facial database as well as images obtained by a webcam and a CCTV camera. The effects of long-distance acquisition are simulated by sub-sampling and rendering out-of-focusing and motion blurs on the original high quality images. The results indicate that our method is effective for single person face detection in the scene at a distance.

The rest of the paper is organized as follows. In Sec. 2, the optimum composite filter with the observation model and other verification stages are described. Experimental and simulation results are presented in Sec. 3. The conclusion follows in Sec. 4.

2. Single person face identification system

Our face identification system is composed of several stages. Figure 1 shows the block diagram of the system.

In the following subsections, each stage will be demonstrated in detail.

In this paper, the optimum composite filtering embraces the observation model during the filter design. A training image in the scene with the long-distance observation model [2,3] is described as follows:

\[ r_t = DBM r_{t-1} + n_t, t = 1, ..., N_t, \]

where \( D \) is a sub-sampling matrix, \( B \) is a blur matrix to model the camera’s point spread function, and \( M \) is a warp matrix; \( r_t \) is the \( t \)-th original training image vector; \( n_t \) denotes an additive noise vector and \( N_t \) is the number of the training images. In the experiment, the warp matrix is set to the identity matrix for the sake of simplicity. It is assumed that one of training images in Eq. (1) can be present in the scene associated with a random binary number:

\[ s(i) = \sum_{t=1}^{N_t} n_t r_t(i - r_t) + n(i), i = 1, ..., N_p, \]

where \( s \) denotes the input scene; \( n \) is the \( t \)-th training image; \( v_t \) is a binary random number which indicates whether or not \( n \) is present in the scene; \( r_t \) is the random location of \( r_t \); \( n \) is assumed to be additive noise with zero mean and \( \sigma^2 \) variance, and \( N_p \) is the number of pixels in the input scene. The probability that \( v_t \) is alternative is \( P(v_t = 1) = 1/N_t \) and null is \( P(v_t = 0) = 1-1/N_t \) [11]. The experiment, we ignore the noise term in Eq. (1) and \( \sigma^2 \) in Eq. (2) is set to 0.01 or 0.1. The output of the correlation filter becomes

\[ o(j) = \sum_{j=0}^{N_t-1} h^*(i+j)s(j), \]

where \( h \) is the complex filter and \( * \) denotes the complex conjugate. The composite filter in the spatial and frequency domain satisfies the following constraints, respectively:

\[ o_t(0) = \sum_{j=0}^{N_t-1} h^*(j)r_t(j) = C_t, \]
\[ \sum_{k=0}^{N_p-1} H_k^*(k)R_k^t(k) = N_p C_t \] (5)

where \( H \) and \( R_k \) denote the composite filter and the \( t \)-th training image in the frequency domain, respectively, and \( C_t \) is an arbitrary constant. With the constraint of Eq. (5), the optimum composite filter minimizes

\[ \frac{1}{N_p} \sum_{k=0}^{N_p-1} |H(k)|^2 E[N(k)] + \frac{1}{N_p} \sum_{k=0}^{N_p-1} |H(k)|^2 S(k)^2 \] (6)

where \( N \) and \( S \) denote the noise and the input scene in the frequency domain, respectively. After the calculation, the optimum nonlinear filter becomes [11]

\[ H(k) = \frac{\sum_{t=1}^{N_t} \lambda_t R_t(k)}{2[N_p\sigma^2 + S(k)^2]} \] (7)

where \( \lambda_t \) is the filter weight, \( t \)-th training image and the input scene. After filtering in the frequency domain, several windows are selected to compose a set of candidate windows. If the following equation is held, a window centred at pixel \( i \) is chosen as a candidate facial region:

\[ \frac{\sigma(i)}{\text{mean}(a)} \geq \varepsilon_f \] (8)

where \( \varepsilon_f \) is a threshold value, which is larger than 1.

B. Skin colour test

The candidate regions are tested to verify the colour of the pixels in the window [13]. A candidate window is accepted if it satisfies more than five conditions in Table 1.

<table>
<thead>
<tr>
<th>Number</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( m_r &gt; \varepsilon_r )</td>
</tr>
<tr>
<td>2</td>
<td>( m_g &gt; \varepsilon_g )</td>
</tr>
<tr>
<td>3</td>
<td>( m_b &gt; \varepsilon_b )</td>
</tr>
<tr>
<td>4</td>
<td>( m_r &gt; m_g )</td>
</tr>
<tr>
<td>5</td>
<td>( m_r &gt; m_b )</td>
</tr>
<tr>
<td>6</td>
<td>( \text{abs}(m_r - m_g) &gt; \varepsilon_d )</td>
</tr>
</tbody>
</table>

Table 1. Skin colour conditions

In Table 1, \( m_r \), \( m_g \) and \( m_b \) denote the average of the red, green, and blue components in the candidate window, respectively, and \( \varepsilon_r, \varepsilon_g, \varepsilon_b \) and \( \varepsilon_d \) are the pre-determined threshold values. They are set to 95, 40, 20 and 15, respectively.

C. Edge mask filtering

The window verified during the skin colour test are subject to the shape test of the face. The following edge mask filtering test is employed [14]:

\[ \sum_{i=1}^{N_w} h_k(i)w(i) > \varepsilon_e \] (9)

where \( h_k \) is an edge mask; \( w(i) \) is the edge image of the window region; \( N_w \) is the number of pixels in the window and \( \varepsilon_e \) is a threshold value. Figure 2 shows the 16×16 pixel edge mask which is used in the experiments.

**Figure 2. Edge mask.**

3. Experimental and simulation results

This section is composed of two subsections. In the first subsection, the composite filter with observation modelling is verified with a facial image database. In the second subsection, images captured by a webcam and a CCTV camera are tested to evaluate the system performance.

A. Improved composite filter with long-distance observation modelling

A facial image database is employed to test the efficiency of applying the observation model. The sub-sampling and blur matrices in Eq. (1) are applied to original training images in order to simulate long-distance acquisition. The CMU facial image database [15] is composed of 14 classes (persons). The image size is 64×64 pixels. The composite filter is trained with three images of the class \( A \), as illustrated in Figure 3. Figures 3(a) and 3(c) show the original and ×0.5 scaled images, respectively. Figures 3(b) and 3(d) show the training images with out-of-focus blur rendered on figures 3(a) and 3(c), respectively. Figure 4 shows test image samples, which are degraded in the same manner as the training images. The performances are compared between three cases: (case 1) original filter without blurring, (case 2) original filter with degraded test images, (case 3) improved filter with degraded test images. Fisher ratio (FR) and discriminant ratio (DR) are employed to evaluate the performance [16]. High numbers of FR and DR indicate the high
discrimination capability between a true reference target \((r)\) and false targets \((s)\). \(FR\) shows the normalized distance between the mean values of the true and the false target outputs by the sum of the variance of each hypothesis. \(DR\) is the ratio between the mean values of the true and the false target outputs. They are defined, respectively, as:

\[
FR(s) = \frac{(m_{tr} - m_{ts})^2}{\sigma_{tr}^2 + \sigma_{ts}^2}, \tag{10}
\]

\[
DR(s) = \frac{m_{tr}}{m_{ts}}, \tag{11}
\]

where \(m_{tr}\) and \(\sigma_{tr}^2\) are the mean and the variance of the outputs for the true target and \(m_{ts}\) and \(\sigma_{ts}^2\) are for the false target. The subscript \(r\) denotes a training class, which is set to class \(A\), and the subscript \(s\) denotes a test class, which is varied from class \(B\) to class \(M\). Figure 4 shows the \(FR\) and \(DR\) plots for three cases. As illustrated in the figures, case (3) outperforms case (2), indicating the effectiveness of the observation modelling.

![Figure 4](image_url)

**Figure 4.** Test image samples with blurs (×0.5 scaled). (a) class A, (b) class B, (c) class C, (d) class D, (e) class E, (f) class F.

---

**Figure 3.** Training images, (a) original, (b) with observation model, (c) ×0.5 scaled, (d) ×0.5 scaled and with observation model.

![Figure 3](image_url)

**Figure 5.** (a) FR (original size), (b) DR (original size), (c) FR (×0.5 scaled), (b) DR (×0.5 scaled), (e) FR (×0.3 scaled), (f) DR (×0.3 scaled).

---

**B. Webcam or CCTV camera images**

In this subsection, several images are captured by a webcam or a CCTV camera. Figure 6 shows a training image set captured by a webcam.
Figure 6. Original training image set.

Figure 7 shows the experimental results of the composite filtering on the high resolution scenes [12]. The size of each input scene is 240×180 pixels. The red boxes are the window regions of the highest peak among the filter outputs.

Figure 7. Face identification with a webcam.

The effect of the long-distance capturing is simulated by blur rendering and the following sub-sampling of the input scene. The out-of-focusing and motion blurs are rendered on the input scenes and the training images. Out-of-focus blurs are rendered by applying circular averaging with a radius size of 10 pixels. Motion blurring involves a filter to approximate the linear motion of a camera by 20 pixels, with an angle of 45° in a counterclockwise direction [17]. The sub-sampling matrix is embraced in the observation model, thus the input scene is resized by 30% after the blur rendering. It is noted that the same blur rendering and sub-sampling are applied to the training images, thus the size of each training image becomes 16×16 pixels, which is the same with the window and the edge mask size. The thresholds $\epsilon_f$ and $\epsilon_r$ are set to 3 and 0.8, respectively.

Figures 8 and 9 are the experimental results with the out-of-focusing and motion blurring, respectively. Figure 8(a) is a high resolution input scene. Figure 8(b) is a degraded image by out-of-focusing blurring. Figure 8(c) is a subsampled image of figure 8(b). Figure 8(d) is the candidate windows that are obtained by the composite filter according to Eq. (8). Figure 8(e) shows the windows which has passed the skin colour test. Figure 8(f) is the edge image of figure 8(c). Figure 8(g) shows the results of the edge mask filtering test performed on the windows in figure 8(e). Figure 8(h) shows the final face region. Figure 9 is the test results of the input scene which is degraded by the motion blurring and sub-sampling.

Figure 8. (a) high-resolution image, (b) out-of-focusing blurring, (c) sub-sampling, (d) composite filter results, (e) skin colour test, (f) edge detection, (g) edge mask filtering, (h) final facial region.
Figure 9. (a) motion blurring, (b) sub-sampling, (c) composite filter results, (d) skin colour test, (e) edge mask filtering, (f) final facial region.

Figure 10. (a) high-resolution image, (b) out-of-focusing blurring, (c) sub-sampling, (d) composite filter results, (e) skin colour test, (f) edge detection, (g) edge mask filtering, (h) final facial region.
Figures 10 and 11 show the results of input scenes capturing several human subjects together. The original high-resolution input scenes are degraded by out-of-focus blurring and sub-sampling in figure 10 and motion blurring and sub-sampling in figure 11.

Another experiment is performed with an input scene captured by a CCTV camera. Since the image is captured around 50 m away indoors, no artificial degradation is performed on the input scene. Figure 12 is an original training image set. The training image is resized by 40%, thus the window size is the same as the previous experiments. The thresholds $\epsilon_f$ and $\epsilon_s$ are set to 9 and 0.5, respectively. Figure 13 shows the results of each stage.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Number of windows detected.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 8</td>
<td>78</td>
</tr>
<tr>
<td>Fig. 9</td>
<td>73</td>
</tr>
<tr>
<td>Fig. 10</td>
<td>47</td>
</tr>
<tr>
<td>Fig. 11</td>
<td>166</td>
</tr>
<tr>
<td>Fig. 13</td>
<td>20</td>
</tr>
</tbody>
</table>
the skin colour and edge mask filtering tests. Both skin colour and edge mask tests are required for figures 8 and 9 to reach the final results. Either the skin colour or edge mask test can derive the final results in figures 10, 11 and 13. There is no missing detection for all cases.

4. Conclusions

This paper addresses the person-specific face identification system at a distance. Several stages are combined in order to improve the system performance via false alarm elimination. The correlation filter has advantages in determining the presence of a person and localizing him in a scene. The long-distance observation model is applied to the training images. Candidate window regions are verified with the skin colour and edge mask filtering tests. The experiments have shown that a person of interest captured at a distance can be detected with the presented method. Further investigations on advanced correlation filters remain for future research.

5. Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2011-0003853 and No. 2012R1A1A2008545). We are also very thankful to In-Su Uhm, Hyung Lee and other undergraduate students for their assistance in the experiments.

6. References