Expert Systems with Applications 38 (2011) 6031–6042

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination

Deng-Yuan Huang [a,*], Wu-Chih Hu [b,1], Sung-Hsiang Chang [a,2]

[a] Department of Electrical Engineering, Dayeh University, 168 University Rd., Dacun, Changhua 515, Taiwan
[b] Department of Computer Science and Information Engineering, National Penghu University of Science and Technology, 300 Liu-Ho Rd., Makung, Penghu 880, Taiwan

ARTICLE INFO

Keywords:
Hand gesture recognition
Gabor filter
Principal component analysis (PCA)
Support vector machine (SVM)

ABSTRACT

In this paper, we present a novel approach for hand gesture recognition based on Gabor filters and support vector machine (SVM) classifiers for environments with varying illumination. The proposed method (1) is robust against varying illumination, which is achieved using an adaptive skin-color model switching method; (2) is insensitive to hand-pose variations, which is achieved using a Gabor filter-based gesture angle estimation and correction method; (3) allows users to wear either a long- or short-sleeve shirt, which is achieved using a method that segments the hand from the forearm. To evaluate the robustness of the proposed method, we created a database of hand gestures in realistic conditions. A recognition rate of 96.1% was achieved using the proposed method. A dynamic gesture recognition system is also presented for real-life conditions. In the proposed system, the recognition results improved from 72.8% to 93.7% when the hand-pose correction module was used, indicating that using the responses of Gabor filters to estimate the hand-pose angle is effective.

1. Introduction

Hand gesture recognition is a challenging topic of research due to the increasing demands for human–computer interactions (HCIs) in recent years. Gesture recognition based on visual perception has many advantages over devices such as mice, keyboards, or electronic gloves. Hand gesture recognition provides users with an intuitive means of directly using their hands to interact with a computer. Hands are deformable objects with many degrees of freedom (DOF), which makes the recognition of hand gestures a difficult task. When gestures are used to define a vocabulary or a symbol in an HCI application, their simplicity is emphasized. Hand gestures can be applied to virtual reality environments, image/video coding, content-based image/video retrieval, and video games.

In order to alleviate computational load and to facilitate the segmentation process in gesture recognition, some constraints such as using simple backgrounds and bright lights, and wearing long-sleeve shirts or electronic gloves are often considered (Caplier, Bonnaud, Malassiotis, & Strintzis, 2004). The hand is assumed to move within acceptably small angles. However, backgrounds are complex and have varying illumination in realistic surroundings, and users often wear short-sleeve shirts. Electronic gloves are frequently adopted to determine various gesture parameters, including the hand's position, angle, and the location of the fingertips. The major problem is that users are forced to wear inconvenient and awkward devices, which leads to more complicated and less natural interactions. Therefore, the present paper develops a method that: (1) is robust to varying illumination, which is achieved using an adaptive skin-color model switching method; (2) is insensitive to hand-pose variations, which is achieved using a Gabor filter-based gesture angle estimation and correction module; (3) allows users to wear either long- or short-sleeve shirts, which is achieved using a method that segments the hand from the forearm.

The rest of this paper is organized as follows. Section 2 summarizes the related work on hand gesture recognition. Section 3 describes the methodology of image preprocessing including the modules of adaptive skin-color model switching, hand-pose angle estimation and correction, and hand-region segmentation. Then, an overview of the principles of Gabor filters for hand representation, the PCA method for dimensionality reduction, and the SVM method for gesture classification are given. The experimental results and analysis are discussed in Section 4. The conclusion is given in Section 5.

2. Related work

In a hand gesture recognition system, hand detection is used to determine the hand's position, scale, and angle. The extraction of

* Corresponding author. Tel.: +886 48511888x2198; fax: +886 48511245.
E-mail addresses: kevin@mail.dyu.edu.tw (D.-Y. Huang), wchu@npu.edu.tw (W.-C. Hu).
[1] Tel.: +886 69264115x3503; fax: +886 69277361.
[2] Tel.: +886 48511888x2198; fax: +886 48511245.

0957-4174/$ - see front matter Crown Copyright © 2010 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.11.016

skin-tone colors is often preceded by hand detection (Conseil, Bourennane, & Martin, 2007; Soriano, Martinkauppi, Huovinen, & Laaksonen, 2000), since it is one of the most important hand features that is invariant to the changes of the hand's pose and scale. However, color-based approaches are not robust in detecting skin-tone colors in the presence of complex background and varying illumination. Thus, an adaptive color space switching method is often applied for surroundings with varying illumination (Chang & Chang, 2006; Stern & Efros, 2005).

Compact skin clusters can be formed in the YCbCr color space under a wide range of lighting variations (Chai & Bouzerdoum, 2000). However, skin color models using the YCbCr color space frequently misclassify non-skin pixels at low luminance as skin-tone pixels and vice versa (Hsu, Abdel-Mottaleb, & Jain, 2002) due to its nonlinear dependence on luma. Since skin-tone pixels have a distinct shape in the normalized color space (r, g), Soriano et al. (2000) used the (r, g) color space to detect skin colors under daylight, incandescent light, fluorescent light, and a combination of these light sources. Under some light conditions, the skin color distribution can be modeled by a Gaussian function in the (r, g) color space (Yang & Waibel, 1996). However, since a single skin color model cannot deal with similar skin-tone pixels in the background and wide ranges of lighting changes, we proposed a method that adaptively switches to an optimal skin-color model for hand detection.

The method of elastic bunch-graph matching (EBGM) is used to classify hand gestures for grayscale images; the method represents the 2D views of hand gestures by labeled graphs (Triesch & von der Malsburg, 1996, 2001). For the EBGM method, graph nodes and edges are labeled with jets and distance vectors, respectively, where jets are vectors based on a 2D Gabor-wavelet transform. In their system, the recognition rates can reach 92.9% and 86.2% under simple and complex backgrounds, respectively. This approach can achieve user-independent and scale-invariant recognition. However, their method is not view-independent because using one graph for one hand gesture is insufficient, and it is computationally inefficient due to the matching process of a single model graph onto an image.

To deal with various angles in finger guessing games, Chen and Tseng (2007) used multiple webcams to construct a multi-angle hand gesture recognition system. In their system, three webcams were set up in front of, to the left of, and to the right of the subject's hand, respectively. Three SVM classifiers were trained for the system. The output classifiers were then used to determine the gesture. The recognition rates for the front, left, and right classifiers were 73.3%, 87.5%, and 92.5%, respectively. However, only three hand gestures were recognized in their work.

Amin and Yan (2007) applied PCA and Gabor filters to recognize the American Sign Language (ASL) finger alphabet from hand gesture images. In their work, they adopted Gabor filters to represent the ASL finger alphabet. The PCA method was then used for features extraction because the dimensionality of Gabor filtered images is very large. The Fuzzy-C-Mean clustering method was then used for gesture classification. A recognition rate of 93.23% was achieved for the ASL alphabet. However, the determination of an optimal combination of principal components, which minimizes the false classification of the Fuzzy-C-Mean method, is quite difficult and time-consuming. Haar-like features and the AdaBoost learning algorithm, which were first used in face detection, were applied to hand gesture recognition by Chen, Georganas, and Petriu (2008). With this method, a recognition accuracy of above 90% and real-time performance can be achieved. However, in their work, only four hand gestures with rotation angles within ±15° against a white-wall background were used.

The Fourier descriptor (FD) is an efficient feature for shape representation in pattern recognition. Conseil et al. (2007) proposed using FD for hand gesture recognition in a vision-based approach. In their work, the authors also provided a performance comparison with Hu moments. The experiments were carried out on both the Triesch hand gesture database (Triesch & von der Malsburg, 1996) and their own, which consisted of 11 gestures performed by 18 individuals with each gesture recorded by 1000 images. Results show that FD gives relatively good recognition rates (84.58%) in comparison with Hu moments (71.08%). However, in their experiment, the requirement of wearing long-sleeve shirts avoids the issue of wrist detection and forearm separation. 500 images for each gesture per person were used in the training phase.

Gabor features have recently been used in the the fields of hand gesture and face recognition (Amin & Yan, 2007; Liu, 2004). The representations of hand gestures based on Gabor features have several advantages, such as localizability, orientation selectivity, and spatial frequency characteristics. However, Gabor features currently adopted by most systems are of too high dimensionality to be used effectively in practical applications. As mentioned earlier, most studies confine their hand gestures to rotations within a small range of angles (Chen et al., 2008), or to wearing long-sleeve shirts (Conseil et al., 2007), or to using simple and bright backgrounds (Chen et al., 2008; Triesch & von der Malsburg, 1996). Thus, we propose a method that deals with these realistic scenarios.

3. System description and related methods

Fig. 1 shows a flow diagram of the proposed hand gesture recognition system. The input video sequence of hand gestures (320×240 pixels) is captured on a Sony PDR-230CAR-A CCD camera (6.0–60 mm F1.4 CCTV lens). The sequence is then sent to the UPMOST UPG301B II image acquisition card for processing. The gesture images are first sent to a preprocessing unit, which consists of the skin-color detection and binarization, hand-pose angle estimation and correction, and hand-region segmentation modules. The detected gesture is then cropped and resized to 20×20 pixels, and convoluted with a Gabor filter, which has five scales and eight orientations. Therefore, 40 filter responses from each image can be obtained. Each filter response is converted into a pattern vector without down-sampling. For instance, for a 20×20 filter response, we have 400 elements in the pattern vector. With 40 filter responses, we obtain a vector with a dimensionality of $400 \times 40 = 16000$ when all of them are concatenated by row (or column). Then, the PCA method is applied to reduce the large dimensionality of the Gabor filtered image. With the reduced Gabor features, the SVM method is used to carry out the hand gesture recognition task. The results are compared with those obtained by metric measures of Euclidean distance and cosine distance.

The Jochen Triesch hand gesture database (Triesch & von der Malsburg, 1996) is available online (http://www.idiap.ch/resources/gestures/). It consists of 10 hand gestures performed by 24 people against three backgrounds. However, using the Jochen Triesch database has some limitations: the number of samples is very small, only gray images are available, and the variability in size and orientation of the hand gestures are relatively small. To overcome these drawbacks and to simulate more real-life conditions, we created a database with 11 gestures (see Fig. 2); this database was inspired by Conseil et al. (2007). The hand gestures were performed by 10 people, most of whom were not familiar with gesture recognition.

The following sections describe how our database is used for training the SVM classifiers. Then, the methods used for image preprocessing, such as the skin-color detection and binarization, hand-pose angle estimation and correction, and hand-region segmentation modules, are briefly reviewed. Finally, the principles of the PCA and SVM methods are given.

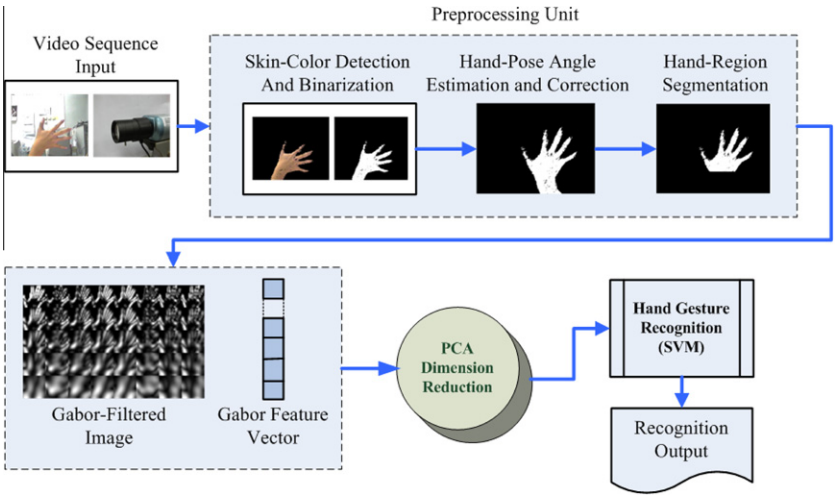

Fig. 1. Flow diagram of proposed video system for hand gesture recognition.

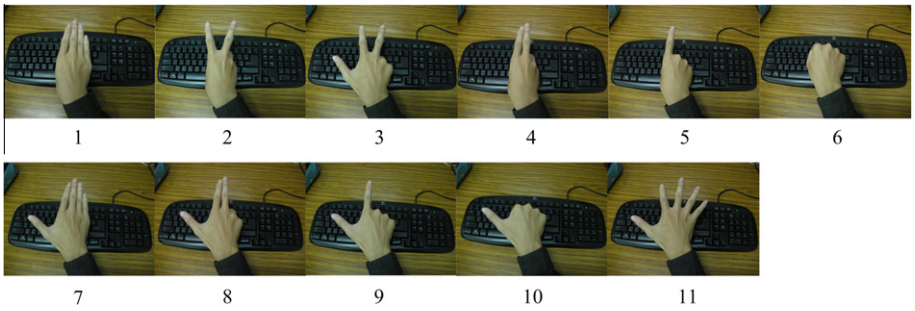

Fig. 2. Eleven hand gestures used in custom dataset.

3.1. Custom database of hand gestures

To establish a hand gesture database, some critical factors such as lighting conditions, hand poses, and the scale variability of gestures must be carefully considered. To train the SVM classifiers, images of 11 hand gestures (see Fig. 2) were collected for 10 people. Each person was requested to perform a certain hand gesture 12 times; each time from a different angle and position (see Fig. 3). Then, the hand gesture was cropped manually and resized to 20×20 pixels. Thus, the dataset contains 120 ($=10 \times 12$) images for each hand gesture and a total of 1320 ($=10 \times 12 \times 11$) images for 11 hand gestures.

3.2. Image preprocessing for gesture recognition

The image preprocessing module was developed to achieve these requirements: (1) robust against lighting changes; (2) insensitive to hand pose variations; and (3) flexible enough for users to wear short-sleeve shirts.

3.2.1. Adaptive skin-color model switching method (ASSM)

Many studies have researched color constancy, but performance has been inadequate. Thus, we present a scheme, called the ASSM method, which can adaptively select an optimal skin-color model for detecting hand gestures under unconstrained illumination conditions.

ASSM is constructed using the possible combinations of three skin color models, i.e., the YCbCr model (Chai & Bouzerdoum, 2000; Hsu et al., 2002), Soriano's model (Soriano et al., 2000), and the Gaussian mixture model (Yang & Waibel, 1996), with three methods of lighting compensation, i.e., reference white (Hsu et al., 2002), modified reference white (Xu, 2007), and gray world (Lam, 2005). In the proposed method, lighting compensation is applied before the skin-color model is used. A flow diagram of skin-color detection for hand gestures is shown in Fig. 4. The captured video images are resized from 320×240 to 80×60 pixels for computational efficiency. Then, a normalized quality measure adaptively selects an optimal skin-color model from the possible combinations.

Fig. 3. Typical images of hand gesture "1" with various pose angles.

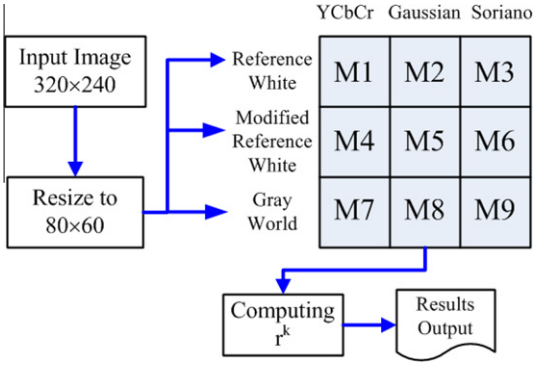

Fig. 4. Flow diagram of skin-color detection for hand gestures using the ASSM method.

3.2.1.1. Reference white lighting compensation. The concept of reference white was first presented by Hsu et al. (2002). In their method, the top 5% of luma values in the image are regarded as the reference white if the number of these pixels is sufficiently large (>100 pixels). The R, G, and B components of a color image are then adjusted so that the average gray value of these reference-white pixels is linearly scaled to 255. Let $i \in [l_u, 255]$ be the top 5% gray levels and f_i be the number of pixels of gray level i in the image. Thus, the modified RGB components can be estimated as:

$$M_{top} = \sum_{i=l_u}^{255} i \cdot f_i \Big/ \sum_{i=l_u}^{255} f_i$$
$$\chi_{new} = \chi_{old}/M_{top} \times 255, \quad \text{where } \chi \in \{R, G, B\} \tag{1}$$

3.2.1.2. Modified reference white lighting compensation. This modified version of reference white (Hsu et al., 2002) was proposed by Xu (2007). In his method, the bottom 5% gray levels are also considered. Let $i \in [l_u, 255]$ and $i \in [0, l_d]$ be the top 5% and bottom 5% gray levels in the image, respectively. The modified (R, G, B) components can be computed as:

$$\chi_{new} = (\ln(\chi_{old}) - \ln(l_d))/(\ln(l_u) - \ln(l_d)) \times 255 \quad \text{where } \chi \in \{R, G, B\} \tag{2}$$

3.2.1.3. Gray world lighting compensation. Gray world (Lam, 2005) is a lighting compensation method that seeks to equalize the mean of the red (R), green (G), and blue (B) channels. The gray world assumption is based on the observation that for a typical scene, the average intensities of the red, green, and blue channels should be equal. Let M and N be the image height and width, respectively. First, the averages of RGB channels and gray levels, χ_{AVG} and μ_{AVG}, are calculated, respectively, as follows:

$$\chi_{AVG} = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} I_\chi(x, y), \quad \text{where } \chi \in [R, G, B]$$
$$\mu_{AVG} = \frac{1}{3}(R_{AVG} + G_{AVG} + B_{AVG}) \tag{3}$$

Then, the scale ratios and modified pixels of the original RGB channels, A_χ and $\widehat{I}_\chi$, are estimated, respectively.

$$A_\chi = \mu_{AVG}/\chi_{AVG}, \quad \text{and} \quad \widehat{I}_\chi(x, y) = A_\chi \cdot I_\chi(x, y) \tag{4}$$

3.2.1.4. YCbCr skin-color model. YCbCr is a family of color spaces (Chai & Bouzerdoum, 2000; Hsu et al., 2002) used as part of the color image pipeline in response to the increasing demand for digital approaches in handling video information; it has become widely used in digital video. In contrast to RGB, the YCbCr color space is luma-independent, which gives it better performance under varying lighting scenes. The corresponding skin cluster used in our system can be described as:

$$\begin{aligned} &(Y, Cb, Cr) \text{ is classified as a skin pixel if :} \\ &\quad 60 \leqslant Y \leqslant 255 \\ &\quad 100 \leqslant Cb \leqslant 125 \\ &\quad 135 < Cr \leqslant 170 \\ &\text{where } Y, Cb, Cr \in [0, 255] \end{aligned} \tag{5}$$

3.2.1.5. Soriano's skin-color model. In this method, a normalized RG color space (r, g) is applied, and then a pair of quadratic functions is used to define the upper and lower bounds of the skin locus, respectively (Soriano et al., 2000). To prevent grayish and whitish pixels from being labeled as skin, pixels are excluded from skin membership if they fall within a circle with a radius of 0.02 around the white point ($r = g = 0.33$). Here, we slightly modified Soriano's skin color model with some RGB constraints, C_{RGB}. Therefore, the skin cluster of value S can be determined with chromaticity (r, g) and the original RGB space as shown in Eq. (6), where the dot $(\cdot)$ means the logical operator "and".

$$\begin{aligned} &S = \begin{cases} 1, & (g < g_u) \cdot (g > g_d) \cdot (R_W > 0.0004) \cdot C_{RGB} \\ 0, & \text{otherwise} \end{cases} \\ &\text{where} \\ &g_u = -1.3767r^2 + 1.0743r + 0.1452 \\ &g_d = -0.776r^2 + 0.5601r + 0.1766 \\ &R_W = (r - 0.33)^2 + (g - 0.33)^2 \\ &C_{RGB} = (R > 130) \cdot (B > 55) \cdot (G > B) \cdot ((R - G) > 25) \end{aligned} \tag{6}$$

3.2.1.6. Gaussian mixture skin-color model. Gaussian mixture models can be viewed as a form of a generalized radial basis function (RBF) network. They are used to describe the skin cluster in the RGB color space as:

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$
$$\text{where } x \in [R, G], 51 \leqslant R \leqslant 102, 51 \leqslant G \leqslant 153 \tag{7}$$

where μ and σ are the mean and standard deviation estimated from the specified skin color regions of R and G channels, respectively. Let $G_{\max} = \mu + 2\sigma$ and $G_{\min} = \mu - 2\sigma$ be the upper and lower bounds of possible skin-tone pixels, respectively. The white point ($R = G = 84$) is also excluded from the skin color membership. Thus, the corresponding skin cluster can be determined as:

$$S(R, G) = \begin{cases} 1, & (G < G_{\max}) \cdot (G > G_{\min}) \cdot (R_W > 26) \\ 0, & \text{otherwise} \end{cases}$$
$$\text{where } R_W = (R - 84)^2 + (G - 84)^2 \tag{8}$$

3.2.1.7. Normalized quality measure. A normalized quality measure, r^{opt}, is presented to optimally choose a skin-color model, denoted as M_k, where $k \in \{1, 2, \ldots, 9\}$, from the possible combinations (see Fig. 4). In the proposed approach, the skin clusters are obtained using a 4-connected component labeling method for each model. In the method, the number of skin clusters with pixels greater than θ_t is computed and denoted as N_C^k, and the total number of the skin-tone pixels in those clusters (with pixels > θ_t) is then accumulated and denoted as N_n^k. As a result, the normalized quality measure for each skin-color model can be estimated as $r^k = N_n^k/N_C^k/(w \times h)$, where w and h represent the width and height of the images, respectively. Thus, the skin-color model can be optimally chosen by r^{opt}, which is defined as:

$$r^{opt} = \arg_{1 \leqslant k \leqslant 9} \max\{r^k\} \tag{9}$$

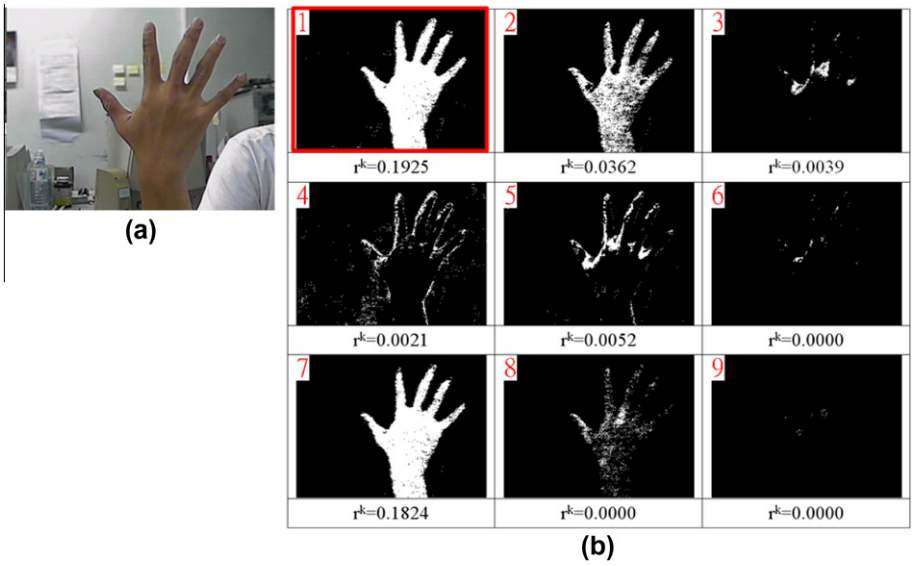

Fig. 5. (a) Original image and (b) skin-tone pixels detected by all possible combinations of skin-color models ($\theta_t = 100$).

Fig. 5 shows the results of detected skin-tone pixels of a hand gesture for all possible combinations with $\theta_t = 100$. An optimal combination of YCbCr + reference white ($r^k = 0.1925$) is selected in this case.

3.2.2. Hand-pose angle estimation and correction

This section describes the proposed Gabor filter-based hand-pose estimation and correction. We first describe the principle of the Gabor filter (wavelet, kernel), and then illustrate how to use the responses of Gabor filters to determine the hand-pose angles. Hand poses can be corrected into an upright orientation using the estimated angle of the hand gesture.

3.2.2.1. Principle of Gabor filters. Gabor filters can capture significant visual properties, such as spatial locality, orientation selectivity, and spatial frequency characteristics. We thus chose Gabor features to represent hand gestures. Mathematically, a 2D isotropic Gabor filter is the product of a 2D Gaussian and a complex exponential function. The general expression is:

$$g_{\theta,\gamma,\sigma}(x,y) = \exp\left(-\frac{x^2+y^2}{\sigma^2}\right)\exp\left(\frac{j\pi}{\lambda}x\cos\theta + y\sin\theta\right) \quad (10)$$

where parameter θ represents the orientation, λ is the wavelength, and σ indicates scale at orthogonal direction. However, with this set of parameters, the Gabor filter does not scale uniformly as the parameter σ changes. We thus use parameter $\gamma = \lambda/\sigma$ to replace λ so that a change in σ corresponds to a true scale change in the Gabor filter. It is convenient to apply a 90° counterclockwise rotation to Eq. (10), such that θ expresses the normal direction to the Gabor wavelet edges. Therefore, the Gabor filter can be alternatively defined as:

$$g_{\theta,\gamma,\sigma}(x,y) = \exp\left(-\frac{x^2+y^2}{\sigma^2}\right)\exp\left(\frac{j\pi}{\gamma\sigma}(x\sin\theta - y\cos\theta)\right) \quad (11)$$

By selectively changing each of the parameters of the Gabor filter, one can tune the filter to a specific pattern in an image. A set of parameters for the Gabor filters used in our system is $\gamma = 0.785$, $\theta = (0°, 90°, 72°, 45°, 36°, -72°, -45°, -36°)$, and $\sigma = (1, 2, 3, 4, 5)$. Some examples of Gabor filters with various parameter values (γ, θ, and σ) and the related coordinates are shown in Fig. 6. However, to reduce computation complexity, note that when the Gabor

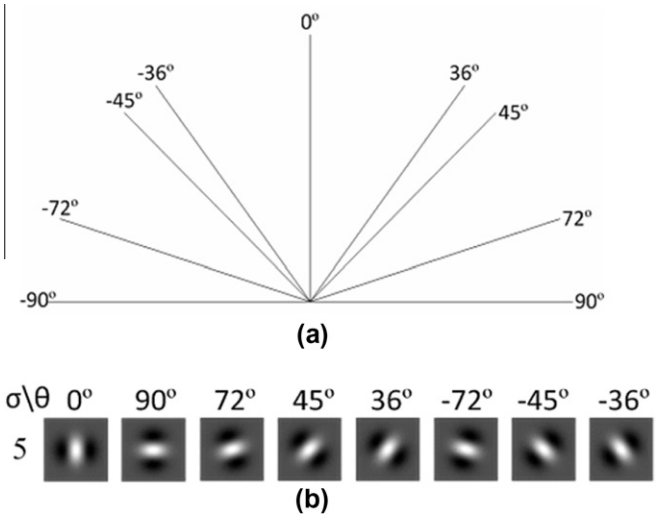

Fig. 6. (a) Coordinates for the orientations of Gabor filters and (b) typical examples of the real part of Gabor filter for parameters $\gamma = 0.785$, $\sigma = 5$, and $\theta = \{0°, 90°, -72°, -45°, 36°, 72°, -45°, -36°\}$.

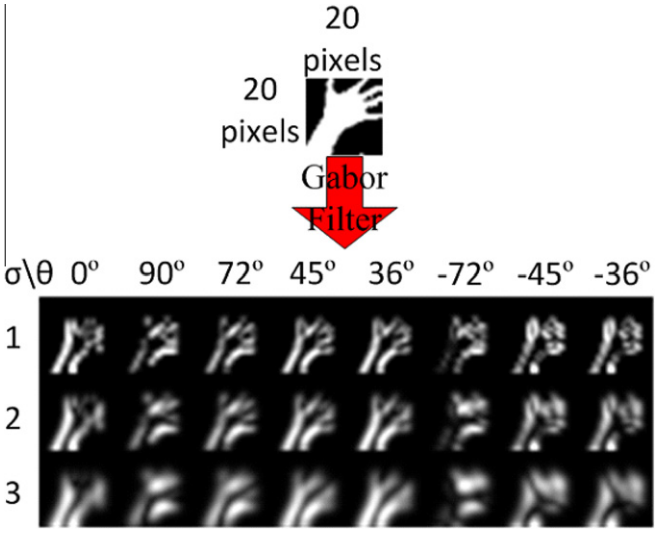

Fig. 7. Gabor filtered image of hand gesture "11" with three scales and eight orientations.

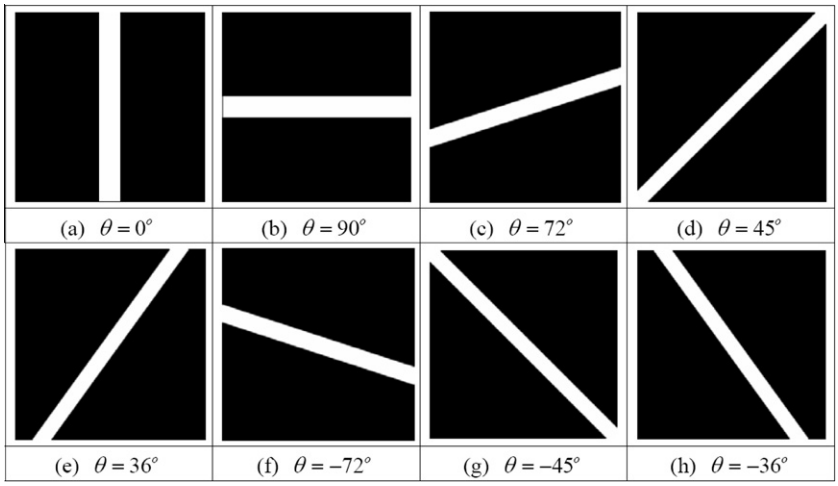

Fig. 8. Images with various angles for white bar for determining hand-pose angle.

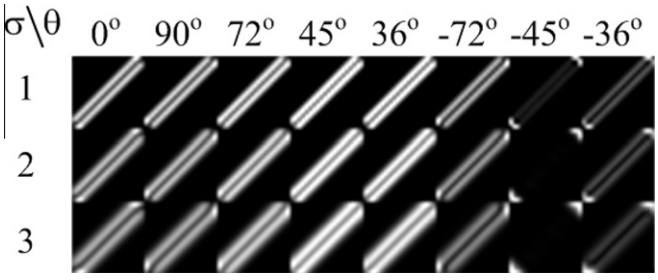

Fig. 9. Gabor filtered images of Fig. 8(d) with an object angle of 45°.

Table 1
Average gray levels computed with all scales ($\sigma = 1, 2, 3$) for Fig. 8(a)–(h) under various orientations $\theta = (0°, 90°, 72°, 45°, 36°, -72°, -45°, -36°)$ of Gabor filters.

	0°	90°	72°	45°	36°	−72°	−45°	−36°
(a) $\theta = 0°$	**176.5**	139.2	169.7	173.2	174.3	169.7	173.2	174.3
(b) $\theta = 90°$	140.5	**176.5**	176.0	173.2	171.8	176.0	173.2	171.8
(c) $\theta = 72°$	115.1	155.9	**162.5**	155.0	150.7	139.8	111.0	99.8
(d) $\theta = 45°$	92.5	92.5	118.7	**162.2**	147.6	58.0	14.5	25.4
(e) $\theta = 36°$	164.7	154.3	170.2	185.8	**186.4**	135.2	87.9	110.6
(f) $\theta = -72°$	115.3	156.2	140.2	111.4	100.2	**162.5**	154.8	150.6
(g) $\theta = -45°$	92.5	92.5	58.0	14.5	25.4	118.7	**162.2**	147.6
(h) $\theta = -36°$	164.7	154.6	135.3	87.6	110.5	170.6	186.0	**186.5**

The bold value means the maximum value in each column.

filters used in the module of hand-pose angle estimation and correction are discussed below, only three scales, i.e., $\sigma = (1, 2, 3)$, are used with other parameters unchanged. The responses of Gabor filters for hand gesture "11" with only 3 scales and 8 orientations are shown in Fig. 7.

3.2.2.2. Hand-pose angle estimation and correction. It is intuitive to use filter responses to determine the hand-pose angles since Gabor filters capture the salient edge and texture orientations. Here, we carried out a simple experiment by placing a white bar against a black background with various angles, as shown in Fig. 8. The filter responses of the selected object angle with 45° (see Fig. 8(d)) are shown in Fig. 9, which indicates that the strongest responses of Gabor filters can be observed near $\theta = 45°$ for all scales. We then estimated the average gray levels with all scales ($\sigma = 1, 2, 3$) for Fig. 8(a)–(h) under various orientations of Gabor filters ($\theta = 0°$, 90°, 72°, 45°, 36°, −72°, − 45°, −36°), as listed in Table 1. The consistency between the object angle and the orientation of Gabor filters with the strongest response suggests that the strongest filter response should be used to estimate the hand-pose angles.

However, in practical applications, the orientation with the strongest response may be different from the actual angle of the hand gesture. To overcome this issue, we first grouped several orientations of the filter responses, and then computed the average gray level for each group, which can be sorted in decreasing order. The group with the greatest response can thus be determined. We can then find the orientation with the greatest response in that group, which is the estimated angle of that gesture. The estimated angle can be used to correct the hand pose into an upright orientation. The details of how to determine the hand-pose angle are presented in Algorithm I (see Fig. 10).

Algorithm I. Hand-pose angle estimation and correction

1. Extract hand gesture region from a binary image, and resizing it to 20×20 pixels.
2. Convolute the obtained image with 24 Gabor filters, and calculate the average gray level for each orientation of filter responses with three selected scales $\sigma = (1, 2, 3)$, i.e., $\gamma^k = \frac{1}{3wh}\sum_{j=1}^{3}\left(\sum_{x=1}^{w}\sum_{y=1}^{h} i(x,y)\right)$, where $k \in \{1, 2, \ldots, 8\}$ corresponds to orientations $\theta \in \{0°, 90°, 72°, 45°, 36°, -72°, -45°, -36°\}$; $i(x,y)$ is the gray level of the Gabor filtered images in the kth orientation; w and h are their width and height, respectively. Note that the average gray levels for ±90° are the same; therefore, only the case of +90° is calculated here.
3. Group the orientations of filter responses, i.e., $g_1 = \{0°, 36°, 45°\}$, $g_2 = \{36°, 45°, 72°\}$, $g_3 = \{45°, 72°, 90°\}$, $g_4 = \{0°, -36°, -45°\}$, $g_5 = \{-36°, -45°, -72°\}$, and $g_6 = \{-45°, -72°, -90°\}$, and calculate the average gray level for each group g_n , i.e., $\varsigma^{g_n} = \frac{1}{3}\sum_{k \in g_n}\gamma^k, n \in \{1, 2, \ldots, 6\}$.
4. Determine the maximum $\varsigma^{g_n^*}_{\max}$, i.e., $\varsigma^{g_n^*}_{\max} = \arg\max_{1 \le n \le 6}\{\varsigma^{g_n}\}$, to find the group g_n^* with the greatest response.
5. Determine the k^*th orientation with the strongest filter responses $\gamma^{k^*}_{\max}$ in group g_n^*, i.e., $\gamma^{k^*}_{\max} = \arg\max_{k \in g_n^*}\{\gamma^k\}$, to estimate the hand-pose angle. For example, if the resulting k^* equals 3, a hand-pose angle of 72° can be obtained.

To evaluate the difference of the known (true) and estimated pose angles, we created a dataset for all hand gestures (i.e., 11 gestures) with known pose angles of $\theta \in \{0°, 36°, 45°, 72°, 90°\}$. For this

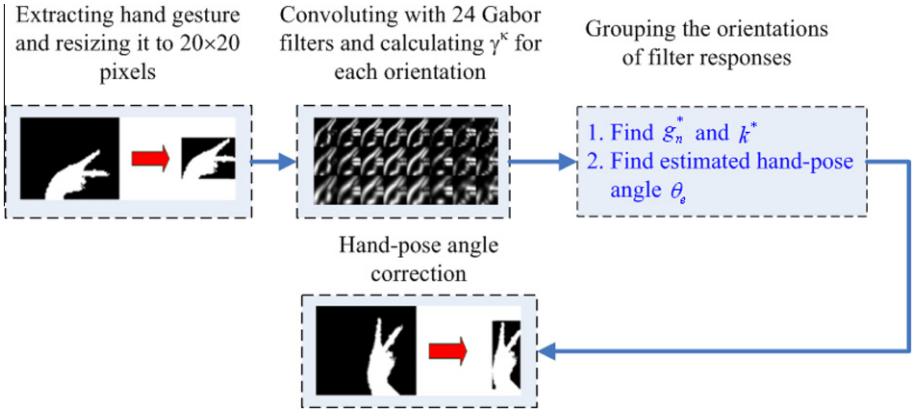

Fig. 10. Flow diagram of hand-pose angle estimation and correction.

dataset, five test samples for each gesture are used; therefore, a total number of 275 (=5 × 11 × 5) samples can be evaluated. Note that only positive hand-pose angles are used here because they allow users to naturally perform the gestures. The difference between the true pose angle, θ_t, and the estimated pose angle, θ_e, for each gesture can be calculated as:

$$\theta_d = \theta_t - \theta_e \tag{12}$$

The difference angle, θ_d, is the resulting orientation of the hand gesture after being rotated by an angle of θ_e in the opposite direction. In general, the smaller the difference angle (θ_d), the higher the recognition rate of gestures that can be achieved, as verified by Amin and Yan (2007). Thus, the percentage of θ_d for all 275 samples within a threshold of θ_m, i.e., $-\theta_m \leqslant \theta_d \leqslant \theta_m$, can be estimated. In addition, if the estimated hand-pose angle is negative, the result will be discarded since users always perform the gestures in the positive direction. With the rule mentioned above, the percentages of θ_d are 49.1%, 73.1%, 96.4%, 96.7%, and 100% for θ_m = 15°, 30°, 45°, 60°, and 75°, respectively. In our dynamic gesture recognition system, the performance of gesture recognition is satisfactory when the angles of the hand gestures are within ±45° (percentage of θ_d is up to 96.4%).

3.2.3. Hand-region segmentation from the forearm

In this work, we used the ASSM method to effectively segment the skin-tone region of hands from complex backgrounds, as shown in Fig. 11. If a user wears a long-sleeve shirt, the skin-tone cluster segmented is exactly the hand region (see Fig. 11(c)); however, when the user wears a short-sleeve shirt, the segmented cluster includes both the hand and arm parts (see Fig. 11(d)),

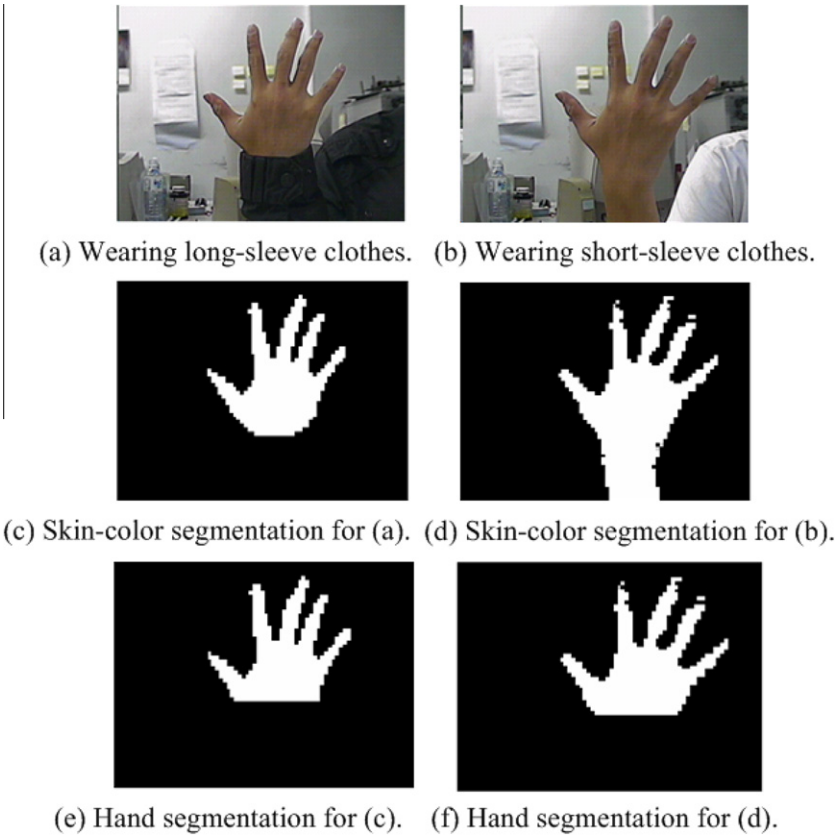

Fig. 11. Hand-region segmentation from forearm for long- or short-sleeve shirts.

Fig. 12. Illustration of hand-region segmentation.

which must be removed. To keep the proposed video system flexible, we present Algorithm II for segmenting the hand from the forearm to eliminate the constraints of wearing only long-sleeve shirts (see Fig. 11(e) and (f)).

Algorithm II. Hand-region segmentation from the forearm

1. Perform row scanning of images with size of $w \times h$ pixels, and record the maximum number of pixels, denoted as α_n (see Fig. 12).
2. Compute the distance from the top of the image to the row labeled α_n, denoted as h^*.
3. Calculate the height, H, of segmented hand region as:

$$H = \begin{cases} H^* + \frac{1}{4}\alpha_n & \text{if } h^* < \frac{4}{5}h \\ h & \text{otherwise} \end{cases}$$

3.3. Feature extraction of hand gestures – the PCA method

The PCA method (Turk & Pentland, 1991) is popular in dimensionality reduction to find a set of orthonormal vectors in the data space, which can maximize the data variance and map the data onto a lower dimensional subspace spanned by those vectors, called principal components. PCA is superior to the linear discriminant analysis (LDA) method, especially for the cases of few training samples used (Martínez & Kak, 2001).

Consider a dataset with M images $x_i \in \Re^N$ $(i = 1, \ldots, M)$ belonging to C subjects, where N (=$w \times h$) is the number of pixels in the image. The total scatter matrix $S_T \in \Re^{N\times N}$ is defined as:

$$S_T = \sum_{i=1}^{M}(x_i - \mu)(x_i - \mu)^T = AA^T \tag{13}$$

where μ is the global mean image of the training set, and $A = [x_1 - \mu \cdots x_M - \mu] \in \Re^{N\times M}$.

A direct computation of S_T is time-consuming due to the large size $(N \times N)$ of the matrix. Instead of directly finding the eigenvector W_{PCA} of S_T, we solve the eigenvalue problem, $RV_{PCA} = V_{PCA}\Lambda$, to obtain the eigenvectors, $V_{PCA} \in \Re^{M\times P}$, and the eigenvalues, $\Lambda = diag[\lambda_1 \cdots \lambda_P] \in \Re^{P\times P}$, with a decreasing order $\lambda_1 \geqslant \cdots \geqslant \lambda_P > 0$, where λ_i represents the nonzero eigenvalues of the matrix $R = A^TA \in \Re^{M\times M}(M \ll N)$. Then, the PCA subspace W_{PCA} is formed by multiplying the matrix A with the eigenvectors V_{PCA}; that is, $W_{PCA} = AV_{PCA} \in \Re^{N\times P}$. Therefore, the feature vector **y** of an image **x** is acquired by projecting **x** into the coordinate system defined by the PCA subspace; that is:

$$y = W_{PCA}^T(x - \mu) \in \Re^P \tag{14}$$

3.4. Classification of hand gestures – the SVM method

Support vectors are samples that lie closest to the decision hyperplane of an SVM, and therefore are the most important data for the determination of the optimum location of the decision hyperplane. In principle, one SVM classifier searches for an optimal hyperplane that maximizes the margins of decision boundaries to ensure that worst-case generalization errors are minimized, which is known as "structural risk minimization (SRM)."

To perform a classification between two classes, a nonlinear SVM classifier is applied by mapping the input data (x_i, y_i) into a higher dimensional feature space using a nonlinear operator $\Phi(x)$, where $x_i \in \Re^d$ and $y_i \in \{+1, -1\}$. Therefore, the optimal hyperplane can be computed as a decision surface:

$$f(x) = \mathrm{sgn}\left(\sum_i y_i\alpha_i K(x_i, x) + b\right) \tag{15}$$

where sgn(·) represents the sign function, and $K(x_i, x) = \Phi(x_i)^T\Phi(x)$ is the predefined kernel function that satisfies Mercer's condition (Vapnik, 1998). In this research, the radial basis function (RBF) is used. It is defined as follows:

$$K(x_i, x) = \exp(-\rho\|x_i - x\|^2), \quad \gamma > 0 \tag{16}$$

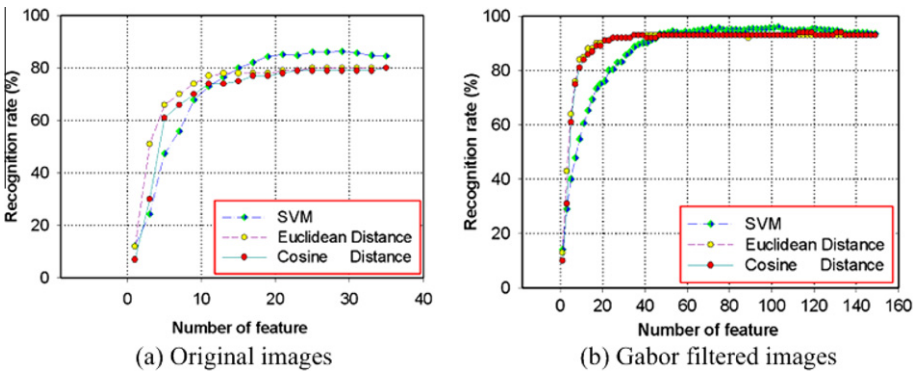

Fig. 13. Recognition rates for the three tested classification methods.

The coefficients α_i and b in (15) can be determined using the following quadratic programming (QP) problem:

$$\begin{aligned} &\max \left[\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right] \\ &\text{s.t.} \quad \sum_i \alpha_i y_i = 0 \\ &\qquad 0 < \alpha_i < C, \ \forall i \end{aligned} \tag{17}$$

The parameter C is a penalty that represents the tradeoff between minimizing the training set error and maximizing the margin. Since the SVM is a binary classifier, it should be extended for an m-class problem in hand gesture recognition. We used the so-called one against one approach, which is a pair-wise method that requires $m(m-1)/2$ SVM classifiers to be trained. In this work, we adopted two sets of parameters of (C, ρ), (3, 0.9) and (2, 0.125), for raw images and Gabor filtered images, respectively, in the training phase for SVM classifiers.

Two similarity measures, i.e., Euclidean distance and cosine distance, are also used for comparison with the SVM method. Let $X \in \Re^d$ and $Y \in \Re^d$ be two pattern vectors. The metrics of Euclidean distance and cosine distance can be defined using Eqs. (18) and (19), respectively. Note that the cosine similarity measure includes

Table 2
Maximum recognition rates for raw images and Gabor filtered images for custom database.

Classification method	Raw images (20 × 20 pixels)		Gabor filtered images (160 × 100 pixels)	
	Recognition rate (%)	No. of features used	Recognition rate (%)	No. of features used
Euclidean distance	80.0	25	93.0	35
Cosine distance	80.0	35	94.0	113
Support vector machine	86.4	29	96.1	103

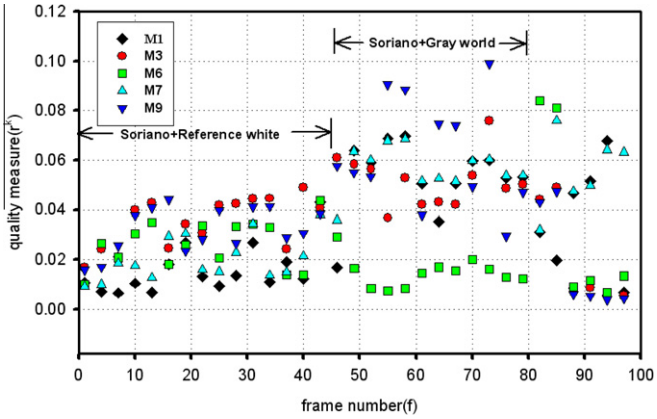

Fig. 15. Variations of quality measure for hand gesture tracking with the adaptive skin color model switching method.

Table 3
Confusion matrix for recognition rates (%) using the SVM method.

	1	2	3	4	5	6	7	8	9	10	11
1	**96.6**	1.7	0	0	0	0	0	0	0	0	1.7
2	0	**100.0**	0	0	0	0	0	0	0	0	0
3	0	1.7	**96.6**	0	0	0	1.7	0	0	0	0
4	5.0	0	0	**88.3**	5.0	0	1.7	0	0	0	0
5	0	1.7	0	1.7	**96.6**	0	0	0	0	0	0
6	0	0	0	0	0	**98.3**	0	0	0	1.7	0
7	0	0	0	0	0	0	**96.6**	3.4	0	0	0
8	0	0	0	0	0	0	3.4	**93.2**	3.4	0	0
9	0	0	0	0	1.7	0	0	6.6	**91.7**	0	0
10	0	0	0	0	0	0	0	1.7	0	**98.3**	0
11	0	0	0	0	0	0	0	0	0	0	**100.0**

The bold value means the maximum value is each column.

Table 4
Comparison results of gesture recognition for various approaches.

	Hand representation	Feature extraction and classification method	Recognition rate (%)	Testing database
Conseil et al. (2007)	Raw image	Fourier descriptor	84.6	11 gestures performed by 18 people, 1000 images per gesture per person
Amin and Yan (2007)	Gabor filtered image	PCA + Fuzzy C mean	93.2	26 sign ASL alphabet performed by 11 people, 12 images per gesture per person
Proposed method	Raw image Gabor filtered image	PCA + SVM	86.4 96.1	11 gestures performed by 10 people, 12 images per gesture per person

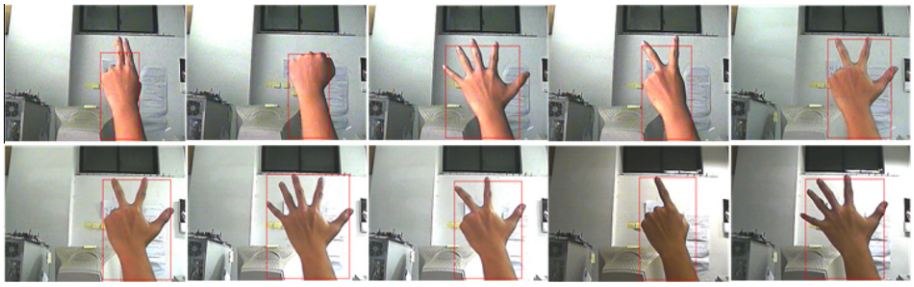

Fig. 14. Hand gesture recognition with varying illumination in a video sequence.

a minus sign in Eq. (19) since the nearest neighbor rule uses minimum distance instead of maximum similarity measure.

$$dist(X,Y) = \sqrt{\sum_{i=1}^{d}(X_i - Y_i)^2} \tag{18}$$

$$dist(X,Y) = -\cos(X,Y) = -\frac{\sum_{i=1}^{d} X_i Y_i}{\sqrt{\left(\sum_{i=1}^{d} X_i^2\right)\left(\sum_{i=1}^{d} Y_i^2\right)}} \tag{19}$$

4. Experimental results and discussion

In this section, we evaluate the performance of recognition using our dataset of gestures. The results are then compared with those reported by Conseil et al. (2007) and Amin and Yan (2007). In addition, the performance of dynamic gesture recognition with the ASSM, the hand-pose angle estimation and correction, and the hand-region segmentation modules is extensively tested.

4.1. Performance evaluation using our database of gestures

The purpose of our database is to widely evaluate the performance of gesture recognition, and particularly to verify invariance to translation, rotation, and scale change for hand gestures. The performance evaluations were carried out on a Pentium PC with a 3.4 GHz processor and 4 GB DDR II RAM memory. The gesture dataset of 1320 images was equally split into a training set and a testing set. For each dataset, 660 images were randomly selected from 6 of 12 samples for each gesture. For SVM classifiers, two sets

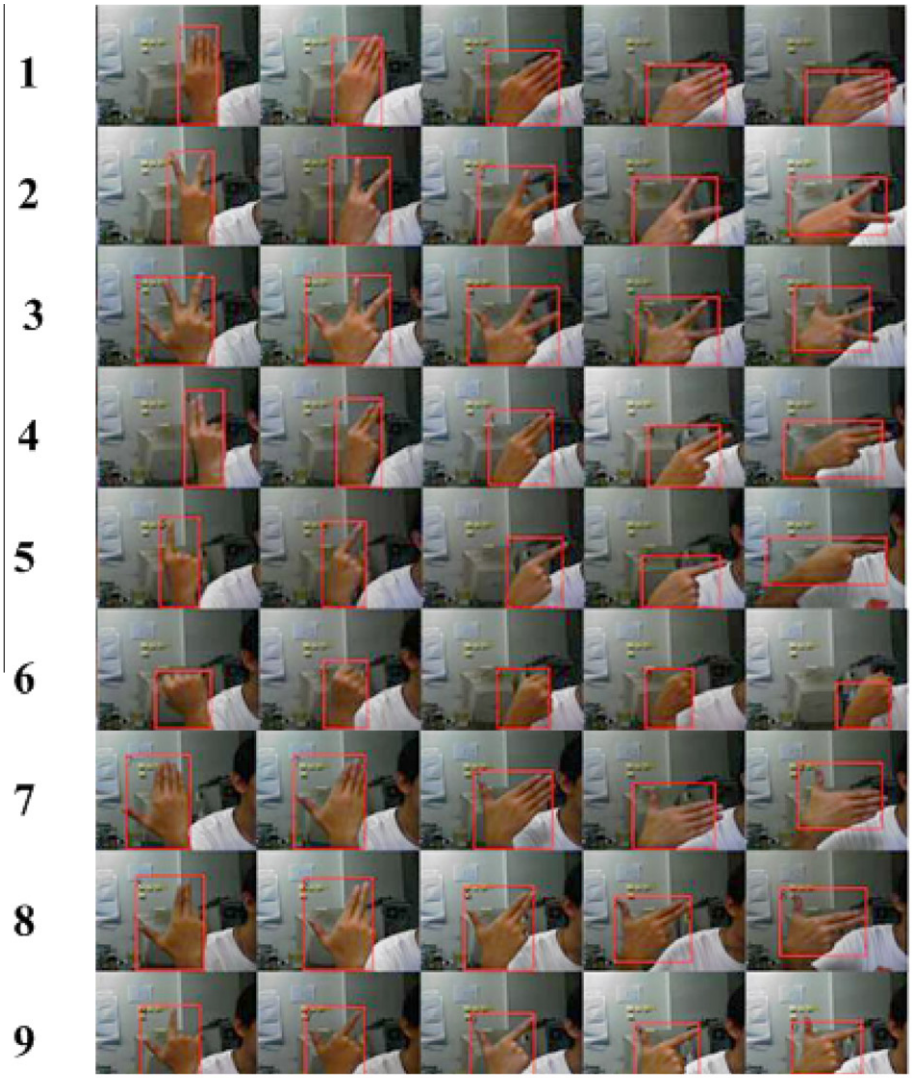

Fig. 16. Results of hand tracking for gestures 1–11 performed with various angles using the SVM method.

Table 5
Comparison results for dynamic gesture recognition with and without the hand-pose angle estimation and correction module.

Type of shirt	Without the hand-pose angle estimation and correction module		With the hand-pose angle estimation and correction module	
	Total number of frames	Recognition rate (%)	Total number of frames	Recognition rate (%)
Short sleeve	260	72.7	270	93.3
Long sleeve	303	72.9	300	94.0
Average		72.8		93.7

of parameters, i.e., $(C, \rho) = (2, 0.125)$ and $(3, 0.9)$, were used for Gabor filtered images and raw images, respectively, to evaluate the performance of Gabor filters.

The recognition rates of hand gestures for raw images with various principal components (features) are shown in Fig. 13(a). The figure shows that the recognition rates are nearly unchanged for the three methods when more than 20 features are used. The maximum recognition rates are 86.4%, 80.0%, and 80.0%, for the SVM method, the Euclidean distance, and the cosine distance, with the features used being 29, 25, and 35, respectively. However, when the Gabor filters are utilized to represent hand gestures, as shown in Fig. 13(b), the maximum recognition rates for the three methods are increased to 96.1%, 93.0%, and 94.0%, respectively. In addition, the representative features are increased to 103, 35, and 113, respectively, due to the larger dimensionality for the Gabor filtered images. As shown in Table 2, the comparative results indicate that the Gabor filters can capture the most discriminating power of hand gestures to be recognized. Moreover, the classification capability for the kernel method (i.e., SVM) is clearly superior to the metric measures of similarity, i.e., the Euclidean and cosine methods. The recognition rates show the effectiveness of using Gabor filters for hand representation and SVM for gesture recognition.

The analysis of the confusion matrix, shown in Table 3, was carried out using the SVM method with 103 features; the analysis indicates which gestures are the sources of error for classification. For instance, gestures 4 and 5 are quite similar in shape from some viewing angles, which lead to a misclassification rate of gesture 4 as gesture 5 of 5.0%. Hand gesture 4 is also erroneously recognized as gesture 1 at a rate of 5.0%. The reason for this may be due to the larger rotation angle of hand gesture 1, making these two gestures look similar at some angles. A similar situation exists for gestures 8 and 9; gesture 9 is misclassified as gesture 8 at a rate of 6.6%. This is not really surprising since gestures 8 and 9 were selected to test the limits of recognition.

Table 4 lists the results of gesture recognition for the proposed method and those reported by Conseil et al. (2007) and Amin and Yan (2007). In the case of only raw images, recognition rates of 84.6% and 86.4% can be achieved by Fourier descriptors (Conseil et al., 2007) and by our proposed method, respectively. However, the results can be improved to 93.2% and 96.1% by using the Gabor filtered-based Fuzzy-C-Mean (Amin & Yan, 2007) and our SVM methods, respectively. The results of Amin and Yan (2007) are satisfactory, but the determination of optimal combinations of principal components is quite difficult and time-consuming.

4.2. Evaluation of adaptive skin-color model switching method

Fig. 14 shows the detection of hand gestures with varying illumination in a video sequence using the ASSM method. To evaluate the robustness of this method, we set up two rows of fluorescent lamps on the ceiling of our laboratory to vary the brightness. For the ASSM method, the input video image (320×240 pixels) is first resized to 80×60 pixels. The quality measures r^k are then estimated by all possible combinations of the skin color models. The maximum r^k indicates the optimal skin color model.

The tracking results of quality measure by ASSM in the video sequence are shown in Fig. 15. Note that for clarity, the results of only five models are shown in this figure. As shown in Fig. 15, the model of M3 (i.e., Soriano + reference white) dominates the first period with a relatively bright condition. When the brightness becomes dimmer, the model of M9 (i.e., Soriano + gray world) is selected. However, when the brightness is further reduced, the models of M1 (i.e., YCbCr + reference white) and M7 (YCbCr + gray world) are alternately used in the relatively dark condition. Clearly, a single skin color model cannot accommodate all the lighting variations.

4.3. Performance evaluation for dynamic gesture recognition

A dynamic gesture recognition system was implemented based on the flow diagram depicted in Fig. 1. In the experiments, the image sequences were extracted by sampling every fourth frame to avoid indistinguishably marginal differences of gesture changes. The results of hand tracking for gestures 1–11 performed at various angles are shown in Fig. 16. The processing time for each frame was 250 ms. Table 5 shows the results of gesture recognition with and without the use of the hand-pose angle estimation and correction module. As shown in this table, the recognition rates are quite close for short- or long-sleeve shirts, indicating that the method used for segmenting the hand from the forearm is effective. Moreover, the average recognition rate is improved from 72.8% to 93.7% when the hand-pose angle estimation and correction module was adopted.

5. Conclusion

In this paper, we have proposed a method for hand gesture recognition that overcomes challenges such as lighting changes, hand-pose variations, and hand detection. The skin-tone clusters of hands can be detected using the method of skin-color model switching, which can adaptively choose an optimal skin-color model as the illumination conditions change. The responses of Gabor filters are used to determine the hand-pose angles. Moreover, the method used for segmenting the hand from the forearm can effectively eliminate the requirement of wearing long-sleeve shirts.

We have also created a realistic database of hand gestures. The performance of gesture recognition using our database is extensively evaluated. In the case of only raw images, the recognition rate of 86.4% is achieved by the proposed PCA + SVM method. The recognition rate is improved from 86.4% to 96.1% when the Gabor filters-based SVM method is used, indicating that the Gabor filters can capture the most discriminating power of hand gestures. In addition, a dynamic gesture recognition system is presented for more real-life conditions. The method for segmenting the hand from the forearm is found to be effective. The recognition result is improved from 72.8% to 93.7% when the hand-pose correction module is applied; this indicates that using the responses of Gabor filters to estimate the hand-pose angle is effective.

Acknowledgement

This research was fully supported by the National Science Council, Taiwan, under Grant NSC-97-2221-E-212-035.

References

Amin, M. A. & Yan, H. (2007). Sign language finger alphabet recognition from Gabor-PCA representation of hand gestures. In *Proceedings of the 6th international conference on machine learning and cybernetics, Hong Kong* (pp. 2218–2223).

Caplier, A., Bonnaud, L., Malassiotis, S., & Strintzis, M. (2004). Comparison of 2D and 3D analysis for automated cued speech gesture recognition. In *Proceedings of the*

9th international conference on speech and computer (SPECOM'2004), St. Petersburg, Russia.

Chai, D., & Bouzerdoum, A. (2000). A Bayesian approach to skin color classification in YCbCr color space. In *Proceedings of TENCON* (Vol. 2, pp. 421-424).

Chang, C. Y., & Chang, H. H. (2006). Adaptive color space switching based approach for face tracking. *Lecture Notes in Computer Science, 4233*, 244–252.

Chen, Q., Georganas, N. D., & Petriu, E. M. (2008). Hand gesture recognition using Haar-like features and a stochastic context-free grammar. *IEEE Transactions on Instrumentation and Measurement, 57*(8), 1562–1571.

Chen, Y. T., & Tseng, K. T. (2007). Multiple-angle hand gesture recognition by fusing SVM classifiers. In *Proceedings of IEEE conference on automation science and engineering, Scottsdale, AZ, USA* (pp. 527–530).

Conseil, S., Bourennane, S., & Martin, L. (2007). Comparison of Fourier descriptors and Hu moments for hand posture recognition. In *Proceedings of the 15th European signal processing conference (EUSIPCO'2007), Poznan, Poland* (pp. 1690-1695).

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

Lam, E. Y. (2005). Combining gray world and Retinex theory for automatic white balance in digital photography. In *Proceedings of the ninth international symposium on consumer electronics, Macau* (pp.134–139).

Liu, C. (2004). Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(5), 572–581.

Martínez, A. M., & Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(2), 228–233.

Soriano, M., Martinkauppi, B., Huovinen, S., & Laaksonen, M. (2000). Skin detection in video under changing illumination conditions. In *Proceedings of the 15th international conference on pattern recognition* (Vol. 1, pp. 839–842).

Stern, H., & Efros, B. (2005). Adaptive color space switching for tracking under varying illumination. *Image and Vision Computing, 23*(3), 353–364.

Triesch, J. & von der Malsburg, C. (1996). Robust classification of hand postures against complex backgrounds, In *Proceedings of the IEEE international conference on automatic face and gesture recognition, Killington, Vermont, USA* (pp. 170–175).

Triesch, J., & von der Malsburg, C. (2001). A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(12), 1449–1453.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: John Wiley and Sons. pp. 423–424.

Xu, J. Y. (2007). *Face detection and recognition technology research in complex background*. M.S. thesis, Shandong University of Technology, China, pp. 22–24.

Yang, J. & Waibel, A. (1996). A real-time face tracker. In *Proceedings of the IEEE workshop on applications of computer vision, Sarasota, Florida, USA* (pp. 142–147).