# Gesture Recognition for Chinese Traffic Police

Fan Guo, Jin Tang, Xile Wang
*School of Information Science and Engineering, Central South University, Changsha 410083, China*
*guofancsu@163.com*

## Abstract

*In this paper, we present a five-part body model to recognize gestures made by Chinese traffic police in complex scenes for driver assistance systems and intelligent vehicles. First, the police's torso and arms are extracted from a complex traffic scene as the foreground region by using dark channel prior and kernel density estimation. Then the coordinates of pixels in the upper arms and forearms are determined using a max-covering scheme, which is based on a key observation that body-part tiles maximally cover the foreground region and satisfy a body plan. Finally, the Gabor feature–based two-dimensional principal component analysis is used to extract the feature of the gestures made by Chinese traffic police. Unlike most previous methods which require a training stage or a 3D measuring device to construct the body part appearance model, we propose to use the max-covering scheme to learn a five-part body model in an automatic way. Experimental results show that good recognition results can be obtained using the proposed method.*

**Keywords:** gesture recognition; Chinese traffic police; five-part body model; Gabor feature; 2DPCA

## 1. Introduction

Gesture recognition of Chinese traffic police has important meaning for driver assistance systems and intelligent vehicles. The goal of the work is to realize the human-machine interaction between the traffic police and the vehicles. However, this is a daunting task since it is very hard to accurately detect traffic police in an unpredictable environment, which makes the problem both complex and limited.

Although gesture recognition of Chinese traffic police is rare in the literature, great progress has been achieved in the recognition of human gesture. Previous gesture recognition methods are mainly divided into two categories: the on-body sensor–based method and the vision sensor–based method. The on-body sensor–based method uses MEMS inertial sensors such as accelerometers and gyroscopes to measure motion and posture. For example, in [1], on-body sensors were fixed on the back of each hand of the police to extract gesture data. Although the method can achieve a good recognition rate, the extra hindrance to the performer and the relatively high cost limit its use in police gesture recognition. Because of its convenience and relatively low cost, the vision sensor–based method has been widely used in gesture recognition. The method commonly follows two steps: The first step involves acquiring the gesture video by using a vision sensor and locating human features, then estimating human poses from these obtained features. The second step is gesture recognition based on the extracted human posture and movement. The vision sensor–based method has achieved both scientific and economic success. For example, Singh et al. used the Radon transform to recognize hand gestures used by air marshals for steering aircraft on the runway [2]. However, a relatively stationary background of video sequence is a must for this method, which is not true for a traffic scene. Kang et al. used upper-body gestures as the interface between a video game and its player and achieved an average success rate of 93.36% for the recognition of 10 gesture commands [3]. Zhen and Huang et al. developed a video-based system for recognizing characters written with a finger [4]. It allows one to enter characters into the computer program by using the movement of a fingertip. Liu et al. [5] presented a systematic framework for recognizing realistic actions from videos "in the wild". Song et al. [6] proposed a unified framework for body and hand tracking, and a multi-signal gesture database is also introduced in their work.

On the other hand, gesture recognition and human body modeling can be closely related problem since acquiring the motion of arms implicitly solves gesture recognition and constructing a good human body model actually ensures a high recognition rate. Although gesture recognition of traffic police has not been the focus of the literature, substantial advances in human body modeling have been reported. Researchers also propose a tree structure model to represent the human body and

reconstruct 3-D human motion poses. The model consists of rigid parts connected by joints [7, 8]. State-of-the-art pose estimation methods [9-11] typically represented the human body as a graphical model composed of 10 major body parts corresponding to the head, torso, and upper and lower limbs. Eichner et al. [12] propose a novel body part appearance models for pictorial structures. Besides, a 3D measuring device–Kinect has also been used to construct human pose model for gesture or gait recognition as in [13-15]. In this paper, to construct the human body model for gesture recognition, in addition to image processing techniques, we use the max-covering scheme to learn a five-part body model in an automatic way.

The remainder of this paper describes our soft biometrics based algorithm in more detail. We begin by pointing out the problem of existing methods for human gesture recognition. In Section 3, we explain the data flow diagram in our system. The detailed procedure to recognize the gestures of traffic police is described in Sections 4 and 5, whereas in Section 6, we illustrate our experimental results. At the end of this paper, we draw our conclusions about this study.

## 2. The problems of existing methods for human pose estimation

So far, there have been many works to construct the human body, but only few researches focus on the gesture recognition of Chinese traffic police. In Refs. [12, 15], a human body modeling method and a 3D measuring device were adopted based on the tree stick structure or the skeleton tracking function, which are commonly used in human pose estimation.

### 2.1. Human poses estimation based on the tree stick structure

For the tree stick structure method [12], the input is an image and a bounding box around the head and shoulders of a person in the image. The output of the algorithm is a set of line segments indicating location, size, and orientation of the body parts, as shown in Figs. 1(b) and 1(c). As can be seen for the four selected frames in Fig. 1, the tree stick model method was not able to locate the arms with satisfying efficiency (Fig. 1(c)). The reason for this is that the prior locations and the appearance transfer mechanism [16] used in the existing method during body-part detection require a training stage, which is hard to satisfy in all kinds of complex traffic environments. We also notice that the error will increase when the segment is far from the root joint in the five-part body model (e.g., the position error of the forearms is bigger than that of the upper arms). Since the police's arms located by the tree stick model method deviate from the ground truth, the estimated rotation angles will deviate from the ground truth accordingly. Therefore, the correct recognition rate for the tree stick method is very low using the Gabor feature–based 2DPCA method. Furthermore, requiring user interaction also limits its use in police gesture recognition, as shown in Fig. 1 (b).

**Figure 1.** The arm location results using tree stick structure method: The images are as follows: (a) original frames, (b) the input of the tree stick structure method, and (c) body-part detection results by the tree stick model method.

### 2.2. Human poses estimation based on the skeleton tracking function

Inspired by researchers' work [13-15], a Kinect device is used to recognize human pose. As a 3D measuring device, Kinect comprises an IR pattern projector and an IR camera. It can output three different images: IR image, RGB image and Depth image. One of the most important functions of the Kinect device is that it can track human skeleton. However, in the experiment, we find that the Kinect method is not suitable for police gesture recognition. There are many reasons for its failure: (1) in a complex traffic environment, it is very likely that traffic police is not the only person in the scene. Thus, it is very hard to determine which skeleton is belong to the traffic police; (2) the Kinect sensor generally has a practical ranging limit of 1.2-3.5m distance, which limits its use in many real applications; and (3) the Kinect device was not able to track the skeleton with high accuracy, especially for the arm location.

Therefore, we can deduce that gesture recognition of Chinese traffic police generally faces two challenges. One is detecting Chinese traffic police in a complex traffic environment. The problem is very hard because of the possibility of high-density crowds and vehicles in the scene. The other challenge involves adopting reasonable human body model and choosing appropriate features for recognizing the traffic police's gestures. To overcome these problems, in this work, we detect traffic police in a complex scene as the foreground region to constrain the arms in a max-covering manner in order to generate a

five-part body model, which we used to determine the relative position and orientation of the arms, and then recognize gestures through Gabor feature–based two-dimensional principal component analysis (2DPCA).

## 3. Overview

The basic idea of our algorithm is to recognize police gestures from the corresponding body parts on the image plane. The positions of the upper arms and forearms in each frame of the video are located with a local search by using the max-covering scheme technique.

The proposed algorithm is divided into three major steps as shown in Fig. 2. The first step is to detect traffic police in a complex scene. The color of reflective traffic vest can be detected using dark channel prior on the police's torso, and the upper arms and forearms of police are also obtained with kernel density estimation (KDE) as part of the foreground region.

In the second step, the upper arms and forearms of traffic police are located using a five-part body model following two steps: (1) obtaining the closed region inside the foreground silhouette on the base of morphological operation and (2) estimating the upper arms and forearms by doing the rotation around the shoulder and elbow joints in a max-covering manner (Fig. 3). Locating arms is the key step of the whole algorithm, and some important human features, such as height and length of arms are used for the arm location step.

In the last step, some typical gestures of traffic police are recognized by using the Gabor feature–based 2DPCA. Here, we use the method to extract the effective features of the traffic police gestures and then adopt a classification method to recognize police gestures in this paper.

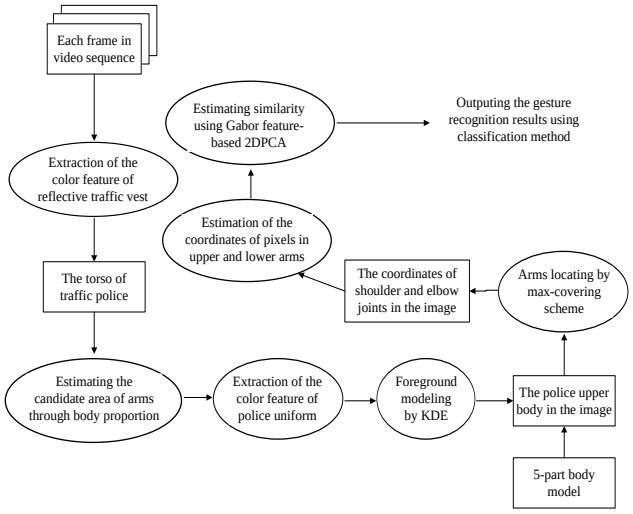



**Figure 2.** Diagram of data flow in our system.

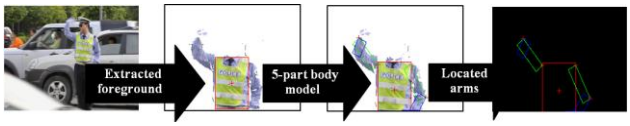



**Figure 3.** Arm location by using a five-part body model.

## 4. Traffic police detection

For the traffic police detection, locating the police is the foundation of analyzing his/her gestures, which can be achieved through two steps: (1) police feature extraction and police foreground modeling, and (2) 5-part body model construction and police arm location.

### 4.1. Police feature extraction and foreground modeling

To the best of our knowledge, the detection of traffic police in a complex traffic environment has not been directly tackled before. The problem involves the possibility of high-density crowds and vehicles in the image. Therefore, we provide a robust solution that makes use of a unique feature of Chinese traffic police, which works efficiently even under such sophisticated scenarios.

According to Chinese regulations, traffic police must wear a reflective vest while they are on duty. Thus, we rely on the soft biometrics–vest to capture the position of traffic police in the image. Two distinctive features of reflective vest are considered here: (1) its apple green color and (2) its strong reflective capacity. The reflective vest can be roughly extracted with color threshold segmentation, which uses the chromaticity coordinates to be more insensitive to small changes in illumination that arise because of shadows. Given three color variables, $R$, $G$, and $B$, the chromaticity coordinates are $r = R / (R + G + B)$, $g = G / (R + G + B)$, and $b = B / (R + G + B)$, where $r + g + b = 1$. We have used the relation between the components. Thus, the color threshold has the following expression:

$$\begin{cases} r - b > 0.01 \\ g - b > 0.17 \\ r - b < g - b \end{cases} \quad (1)$$

Thus, for example, the vest, the tree, and the plant are green. Because of that, these objects were extracted by color thresholding. Note that some false detection is not easy to eliminate by only using color threshold segmentation because the operation might also extract other objects with a similar color. Therefore, reflective capacity as another important feature is considered here. We notice that the intensities of the reflective vest in three color channels, $R$, $G$, and $B$, all have very high values because of the strong reflective capacity, whereas for other colorful objects or surfaces (e.g., the green grass, the tree, the plant), at least one color channel has very low

intensity in some pixels. Thus, the dark channel prior [17], which was proposed to solve the dehazing problem, is used here to further extract the vest. Formally, for an image *J*, its dark channel $J^{dark}$ is defined by

$$J^{dark}(x) = \min_{c\in\{r,g,b\}} (\min_{y\in\Omega(x)} (J^c(y))), \quad (2)$$

where $J^c$ is a color channel of *J* and $\Omega(x)$ is a local patch centered at $x$. A detected pixel, $x$, will be considered to be part of the reflective vest only if $J^{dark}(x) > T$, as shown in Fig. 4(b). The value of *T* is application based. We have decided to keep it fixed at 85 for all results reported in this paper. Fig. 4(c) shows the result in the red bounding box. Note that if the height of the bounding box is too small (e.g., less than 1/20 of the image height), which means no police is detected, then there is no need to do the following steps.

(a) (b) (c)

**Figure 4.** Reflective vest extraction.

For many researchers, the human body can be regarded as an example of perfect proportions. According to their theory, a perfect body is eight heads high. The shoulder is two head lengths wide. The upper arm is one and a half heads long, and the forearm is one and a quarter heads long. Thus, it can be deduced that the whole arm is about three heads long. We use these proportions to narrow the search area for the arms of traffic police. The proportion of the human upper body and the search area is depicted in Fig. 5.

**Figure 5.** The proportion of the human upper body.

For the purposes of discussion, we define the torso and the arms of traffic police in the scene as the foreground and other parts of the scene as the background. The search area covers the possible positions of the arms. However, the background regions contained in the area do not provide any information about the arms. In fact, the background context causes ambiguity, which eventually results in false body model construction. In our experiments we found that it is better to estimate one distribution for the background and one distribution for the foreground using a kernel density estimator [18]. Assuming that the police's torso will be centered in the red bounding box (Fig. 6(a)), we first extract the pixels that satisfy the color threshold constraints: $b - g > 0.05$ and $b - r > 0.05$, as shown in Fig. 6(b). All the pixels that belong to the blue color that appear outside the bounding box are sought in the search area. Then as shown in Fig. 6(c), we use the top 20 pixels that are close to the center point of the box as the samples to estimate the foreground PDF. Let $x_1, x_2, ..., x_{20}$ be a sample of intensity values for a pixel. Given the intensity of target pixel $x_t$, we can estimate the density as

$$\Pr(x_t) = \frac{1}{20}\sum_{i=1}^{20}\prod_{j=1}^{3}\frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2}\frac{(x_{ij}-x_{tj})^2}{\sigma_j^2}}. \quad (3)$$

In Eq. 3, $\sigma_j$ is a suitable bandwidth for *R*, *G*, and *B* three-color channel. In our experiment, $\sigma_j$ is set to $2/0.68\sqrt{2}$. Consequently, we compute pixel probabilities for the foreground and assign every pixel outside the bounding box to its most probable distribution. An illustrative example is shown in Fig. 6(d).

(a) (b) (c) (d)

**Figure 6.** Estimated police foreground image.

## 4.2. Body model and location of the arms

Given the police foreground image, the location of the arms can be simulated as a jigsaw puzzle problem. Here, we represent traffic police by using a five-part body, which is inspired by the widely used 10-part body model. The upper body is our focus. Thus, our five-part body only includes the torso, upper arms, and forearms. Each body part is represented as a rectangle. The five-part body model and the tree structure of the five-part body model are shown in Fig. 7. Notice that the body model consists of rigid parts connected by joints, in which $J_1$ is the root joint corresponding to the clavicle. Information about other joints is provided in Table 1. Fig. 7 shows that the basic body plan follows a tree structure. A local coordinate system is attached to each body part. The orientation of the local coordinate system is also shown in Fig. 7, and the origin of the coordinates is located at the position of each shoulder or elbow joint.

We use the bottom-up method to locate the arms. For this method, body-part candidates are first detected and then assembled to fit the image observations and a body plan. In our proposed method, we first locate the potential torso in target images so that we can use it in the max-covering scheme. Then we use simple box detectors to find arm candidates. Since we have a rough foreground image, the arm candidates can be pruned; we only keep

the candidates that completely cover the foreground pixels. Here, an arm candidate is represented as a rectangle with a start side and an end side.

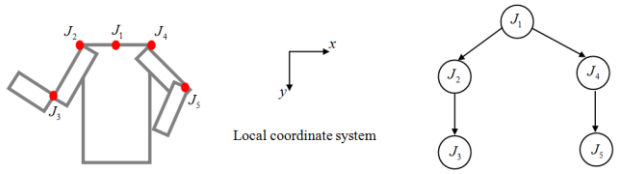



**Figure 7.** Left: five-part body model; right: the tree structure of the five-part body model.

**Table 1.** Information related to the joints of the five-part body model

| **ID** | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ |
|---|---|---|---|---|---|
| Joint | Clavicle | Left shoulder | Left elbow | Right shoulder | Right elbow |

Each arm candidate covers some pixels in the foreground image. Intuitively, the arm tiles should cover foreground pixels as much as possible. Thus, locating the arm is performed with local research in the foreground image by using the length of arms and a max-covering scheme. The whole location process is broken down into two steps: (1) estimating a closed region inside the foreground silhouette based on a morphological operation and (2) locating the position of the upper arms and forearms by rotating around the shoulder and elbow joints in a max-covering manner. These steps are explained in detail as follows.

*Step* 1: Obtaining the closed region inside the foreground silhouette. Once the foreground image is obtained as we explained in Sections 4.1, the max-covering scheme can be formulated as the following optimization problem:

$$\phi = f(\theta, s, r), \tag{4}$$

where $\phi$ is coverage rate, a floating point number from 0 to 1 related to each image pixel. The higher the rate is, the more likely the pixel belongs to the arm. The three parameters $\theta$, $s$, and $r$ are used to control the coverage rate. The value of $\theta$ controls the rotation angle of the joints. The parameters $s$ and $r$, respectively, specify the length and the width of each arm represented by a rectangle. A typical value of $r$ is 1/6 the length of the torso, according to the human body proportion. We adjust the value of $\theta$ and $s$ to control the number of pixels that are covered in the foreground. This optimization, thus, tends to find the position of the rectangle that makes $\phi$ reach its maximum value, 1, which means that the rectangle completely covers the foreground pixels by using the proposed method. The max-covering scheme in Eq. 4 is a local search problem. We need to find an arm configuration to make $\phi$ equal 1 while satisfying the body plan. It is generally NP-hard because of the incomplete extraction of the foreground introduced by the KDE. We need to make sure that the region inside the foreground silhouette is closed without holes so that it can be completely covered by the variable rectangles with a different $\theta$ and $s$. Thus, a morphological operation is used here to tackle this problem. An illustrative example is shown in Fig. 8.

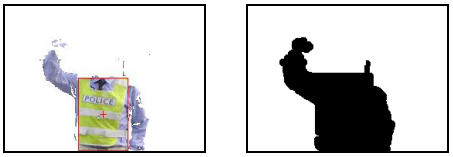


**Figure 8.** Foreground image and its closed binary image.

*Step* 2: Locating the position of the upper arms and forearms. Assume that we have three coordinate systems: the torso, the arm, and image coordinates. In the coordinate system of the left upper arm, the position of a pixel, $P_L$, is given by its coordinates, $(s_1, r_1)$. As shown in Fig. 9(a), the coordinate transformation between the left upper arm plane and the torso plane can be calculated as

$$\begin{cases} x = r_1 \cos\theta + (s - s_1)\sin\theta \\ y = (s - s_1)\cos\theta - r_1 \sin\theta \end{cases}. \tag{5}$$

Let $(u, v)$ denote the coordinates of $P_L$ in the image plane, and we can now translate the coordinates $(x, y)$ from the torso plane to the image plane using

$$\begin{cases} u = x - (r_1 \cos\theta + (s - s_1)\sin\theta) \\ v = y - ((s - s_1)\cos\theta - r_1 \sin\theta) \end{cases}, \tag{6}$$

where $\theta$ is the rotation angle of the left shoulder joint. The variable $u$ is the vertical position of $P_L$, and $v$ is its horizontal position in the image coordinate system. Similarly, as shown in Fig. 9(b), the image coordinate system of the pixel $P_R$ that belongs to the right upper arm can be expressed as

$$\begin{cases} u = x + (r_1 \cos\theta + (s - s_1)\sin\theta) \\ v = y + (r_1 \sin\theta - (s - s_1)\cos\theta) \end{cases}. \tag{7}$$

Therefore, the variable rectangles of upper arms can be obtained by adjusting different $\theta$ and $s$ to completely cover the foreground pixels. Since the positions of arms are estimated in depth-first order as shown in the tree structure of Fig. 7, the detection of the upper arm is then used to guide the search for the forearm. From the estimated elbow joint position, a certain rotation angle is found based on foreground, and the forearm rectangles are converged to local maximums. Eq. 6 and Eq. 7 are also used to obtain the variable rectangles of the forearms.

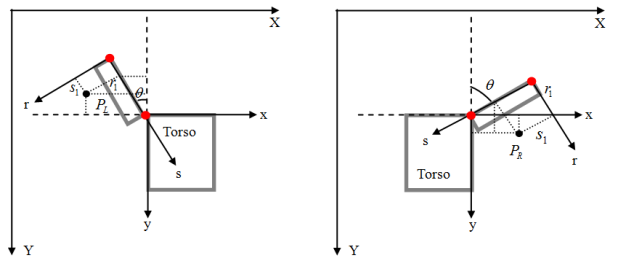



**Figure 9.** Coordinate system of the left arm and the right arm. (a) The coordinate system of the left arm. (b) The coordinate system of the right arm.

## 5. Gesture recognition

The gesture system of the Chinese traffic is defined and regulated by the Chinese Ministry of Public Security. Currently, eight gestures for traffic guidance are included: (1) stop, (2) move straight, (3) left turn, (4) left turn waiting, (5) right turn, (6) change lane, (7) slow down, and (8) pull over. Fig. 10 shows the eight gestures. In our previous work, the rotation angles of the shoulder and elbow joints of each frame are used to match the defined rotation angles of standard gestures [19]. In this paper, we propose to use Gabor feature-based 2DPCA to extract the effective features of the traffic police gestures and calculate the shortest Euclidean distance with the highest degree of similarity from the result of the arms' detection to recognize police gestures.

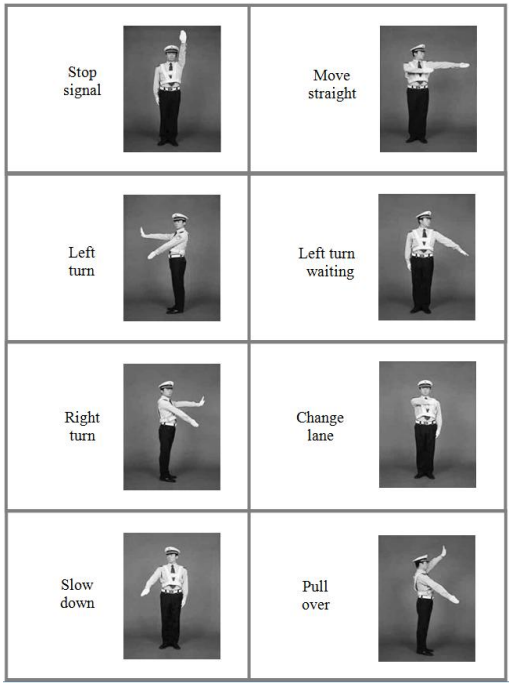



**Figure 10.** Chinese traffic police gestures.

### 5.1. Gabor feature–based 2DPCA

Once the police's arms are located, a method of Gabor feature–based 2DPCA [20, 21], which is used for palm print or face recognition and proved to be one of the best algorithms for object recognition, is adopted to accomplish the gesture recognition task with appropriate parameters. The algorithm consists of three steps: (1) Gabor features of different scales and orientations are extracted through the convolution of the Gabor filter bank and the arm location images, (2) 2DPCA is then applied for dimensionality reduction of the feature space in both row and column direction, and (3) Euclidean distance and the nearest neighbor classifier are finally used for classification. 2-D Gabor has the following general form:

$$G(x,y,\theta,u,\sigma_g)=\frac{1}{2\pi{\sigma_g}^2}\exp\{-\frac{x^2+y^2}{2{\sigma_g}^2}\} \times\exp\{2\pi i(ux\cos\theta+uy\sin\theta)\} \quad , \qquad (8)$$

where $i=\sqrt{-1}$, $u$ is the frequency of the sinusoidal wave, and $\theta$ is the standard deviation of the Gaussian envelope. $\sigma_g$ is a constant, which defines the size of the Gaussian envelope. In our experiment, we set $\sigma_g=2\pi$. In order to extract more effective features on various orientations and scales, a Gabor filter bank is used in our method since the filter has been found to be particularly appropriate for texture representation and discrimination. Thus, Gabor filter can achieve better performance for police gesture recognition compared with other feature extraction methods, such as Hu invariant moments [22], Zernik invariant moments [23], and others.

For Gabor filter, different orientations and scales can extract different image features and thus can obtain different recognition rates. In our experiment, we find that the highest value of recognition rate is achieved when six scales and six orientations are chosen for different local features. This process can be written as

$$u=0.2592/\sqrt{2}^{v} \qquad v=0,1,...,5\,. \qquad (9)$$

$$\theta_k=\frac{\pi(k-1)}{6} \qquad v=1,2,...,6. \qquad (10)$$

Thus, 36 Gabor filters are selected for feature extraction.

### 5.2. Gesture feature extraction using Gabor feature-based 2DPCA

Suppose that there are $N$ 60 × 90 training police arm location images denoted by matrices as $A_i(i=1,2,...,N)$. The convolution of the Gabor filter bank and image $A_i$ yields Gabor feature matrices $H_i(v,k)$ ($v$ = 0,…,5; $k$ = 1,…,6).

For simplicity, we downsample the Gabor feature with a downsampling rate of $\rho = 3$. By concatenating all the 20 × 30 downsampled Gabor feature matrices, $O_i(v,k)$ $(v = 0,...,5;\ k = 1,...,6)$, in the column direction, the Gabor feature matrix $X_i$ of image $A_i$ can be represented as

$$X_i = \{O_i(0,1)^{'}, O_i(0,2)^{'}, ..., O_i(5,6)^{'}\}. \quad (11)$$

The Gabor feature space $X$ is constructed by all the Gabor feature matrices of training samples in the row direction $X = \{X_1, X_2, ..., X_N\}$, the dimension of which is $20 \times 30 \times 36\ N$. If we directly adopt the Gabor features to match the templates, the dimension of image space is very high, which requires too much time and memory. Thus, 2DPCA is used here to effectively reduce the dimension. In 2DPCA, the covariance matrix $G$ can be evaluated by

$$G = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^T(X_i - \bar{X}), \quad (12)$$

where

$$\bar{X} = \frac{1}{N}\sum_{i} X_i. \quad (13)$$

Since the size of $X_i$ is $20 \times 30 \times 36 = 21{,}600$, $G$ has a dimension of $360 \times 60$. The orthonormal eigenvectors of $G$ corresponding to the $d$ largest optimal value are proven to be the optimal projection matrix

$$R = [r_1, ..., r_d]. \quad (14)$$

The value of $d$ can be determined by the ratio of the sum of the chosen $d$ largest eigenvalues to all. In our experiment, we set $d = 10$. That is because the recognition rate is the highest when $d$ is 10. Thus, the dimensions of $R$ are $360 \times 10$, and the ultimate dimension of a Gabor feature vector is reduced from $36 \times 600 = 21{,}600$ to $360 \times 10 = 3{,}600$.

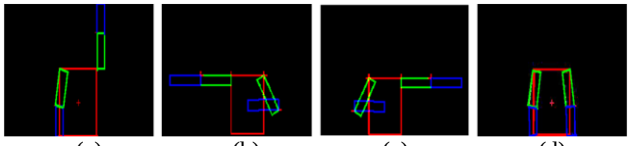
### 5.3. Gesture classification for Chinese traffic police

A nearest-neighbor (NN) classifier is adopted for classification. Supposed that traffic police gesture category $c_i$ ($i$ = 1, 2, …, $p$) has $N_i$ template training samples $B_j^{(i)} = (R_j^{(i)}, R_j^{(i)}, ...R_m^{(i)})$, $(j = 1,2,...,N_i)$ $N = \sum_{i=1}^{p} N_i$ is the total number of training samples in the template database, and these samples are assigned $c_p$ categories, $p$ is the number of traffic police gesture categories used to classify. Supposed that the feature of testing image $B$ would be recognized, Euclidean distance $D_i(B, B_j^{(i)})$ can be computed as the following expression:

$$D_i(B, B_j^{(i)}) = \sqrt{(B - B_j^{(i)})^{\mathrm{T}}(B - B_j^{(i)})} \quad (15)$$

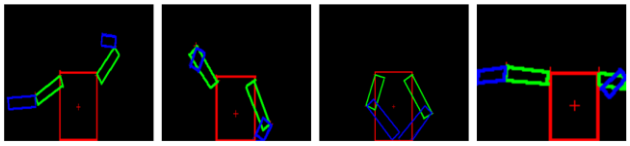
The distance is computed to measure the similarity between $B$ and template samples $B_j^{(i)}$. The following decision rules are taken to judge two things. One is to what categories of Chinese traffic police gesture the recognized gesture belong; the other is whether the gesture is a traffic police gesture. If $D_m(B) = \min_{i \in (1,2,...,p)} \{D_i(B, B_j^{(i)})\}$ and $D_m(B) \le T$, then $B \in c_m$, else $B$ is not a traffic police gesture. $T$ is called similarity threshold.

The arm location results of standard police gestures with different distance and angles constitute the template traffic gesture database. Fig. 11 shows some sample templates for the gestures "stop," "move straight (left/right)," and "no sign." For nearest-neighbor method, a simple illustrative example is shown in Fig. 12. As can be seen in the figure, four arm location results obtained by real captured photos are randomly chosen for testing. From the Euclidean distance obtained by the nearest-neighbor method, we can arrange the standard "move straight (left)" (see Fig. 11(b)) and four testing images in decreasing order of similarity: Figs. 12(d), 12a), 12(b), and 12(c). Thus, we deduce that Fig. 12(d) indicates the "move straight (left)" gesture. This confirms our observation in real captured images.



(a) (b) (c) (d)

**Figure 11.** Template examples: (a) arm location image of "stop," (b) arm location image of "move straight (left)," (c) arm location image of "move straight (right)," and (d) arm location image of "no sign."



(a) (b) (c) (d)

**Figure 12.** Arm location examples with different Euclidean distances (D): (a) D = 1.5931, (b) D = 1.7038, (c) D = 1.8143, and (d) D = 1.2933.

## 6. Experiments and results

We perform the gesture recognition algorithm by executing MATLAB on a PC with a 3.00 GHz Intel Pentium Dual-Core Processor. In our experiment, we not only test for a single image but also extend to videos.

Although the final goal of our project is serving the driver assistance system, as a preliminary analysis of the problem, we mainly focus on the algorithm design at present. To test the proposed contribution, we measure gesture recognition of the Chinese traffic police while their arms in the image were automatically marked by our proposed method. The assumption the proposed algorithm makes is that traffic police are seen approximately from a frontal viewpoint.

### 6.1. Dataset

The real-captured image dataset was recorded in a controlled setting with no camera motion and "clean" background. The dataset contains nine categories of gestures: stop, move straight, left turn, left turn waiting, right turn, change lane, slow down, pull over, and mixed gestures. There were 9 actors, including 1 police officer and 8 students, performing each gestures three times, resulting in about 73 video sequences in total.

Since the real-captured image dataset is relatively simple, the actors can be easily detected. We collected a more complex and challenging dataset based on YouTube videos. The dataset has the following properties: 1) a mix of statistical camera and moving cameras, 2) complex background; 3) variation in object scale, 4) varied viewpoint, and 5) low resolution. This dataset contains 5 video sequences in total. The frame resolution of each video sequence is 320 × 240, and the gestures in these sequences include "stop," "move straight (leftward or rightward)," and six other gestures. Each gesture appears alternately in short intervals of time. It should be stated that the data set is difficult as the scenarios are very complex with high-density vehicles and crowds. We believe that the experimental results on this dataset will be very valuable considering that most previous research experiments were conducted within human-controlled setting to certain degree.

### 6.2. Arm Location Experiments on Real-captured Image Dataset

Since the real-captured image dataset is relatively "clean", the detection of traffic police is much easier in this case. For this dataset, the objective of the arm location experiment is to demonstrate the benefit of the proposed max-covering scheme. In Fig. 13, the images in the second row indicate the marked torso location with the highest precision. Besides, by changing the color segmentation conditions shown in Eq. (1), the forearm rectangles are also displayed even the traffic police wears a short-sleeved shirt. Therefore, we can draw the conclusion that the proposed method to recognize gestures of Chinese traffic police can present encouraging recognition results, as shown in Fig. 13.

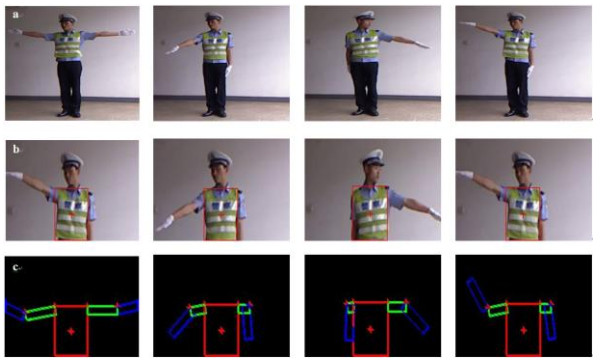


**Figure 13.** Arm locating results for the proposed method. The images are as follows: (a) original frames, (b) the results of detecting the police's torso, and (c) the results of the corresponding arm location.

### 6.3. Arm Location Experiments on YouTube Dataset

We thoroughly tested our proposed method over several challenging video sequences from YouTube dataset. In the following, we evaluate the performance of the proposed algorithm based on 634 frames of traffic police video material. Here, the five-part body model is used to locate both the upper arms and forearms and to prove the efficiency of the proposed method. As can be seen in Fig. 14, challenging test sequences were used to test the effectiveness and accuracy of the method. Experimental results show that the proposed method can obtain precise arm location results for Chinese traffic police, as shown in Fig. 14.

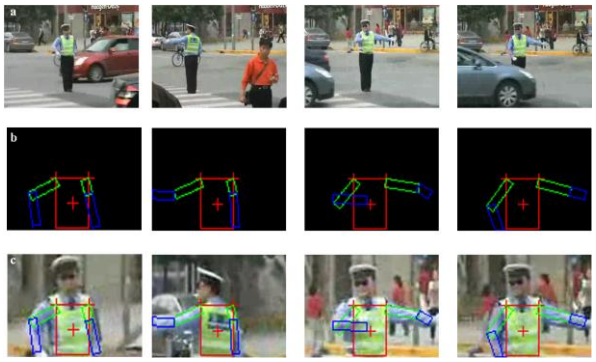


**Figure 14.** Arm locating results for the proposed method. The images are as follows: (a) original frames, (b) and (c) body-part detection results by the proposed method.

For video gesture recognition, transitional gestures are not considered; only the final standard gestures are considered as recognition results. Besides, an image may not contain a policeman or a gesture. In this case, no result is produced, and the intelligent vehicle will keep its current state of movement. In order to reduce the probability of misjudgment as much as possible, a single frame is far from enough, so a fusion strategy should be

designed to make full use of the dynamic characteristics of the video sequence. Here, we defined that the intelligent vehicle will change its state of movement only if the suddenly changing results are the same across three consecutive frames. Otherwise, the vehicle will still keep its current state of movement.

### 6.4. Gesture Recognition Experiments with Gabor feature–based 2DPCA

In this experiment, the correct detection rate of the arm location for police officer and the correct recognition rate of the proposed Gabor feature–based 2DPCA (GB2DPCA) are investigated. The parameter setting of the Gabor filter is depicted in subsection 5.2, and the downsampling rate, $\rho$, remains 3 for the Gabor feature–based algorithm.

We collect eight classes of arm location results of standard traffic police gestures according to Chinese regulations for our template database, and each class has five sample images. All the original images of the samples are captured approximately from a frontal viewpoint. Each frame of the input video sequences can be regarded as the testing image of gesture recognition. Fig. 15 presented the correct detection rate of arm location for police officer and the correct rate of GB2DPCA. As can be seen, all eight gestures can be recognized using this method. Furthermore, there is a "no sign" class (no. 0) in Fig. 15. Otherwise, each time a police is spotted, the closest gesture class will be assumed. In the experiment, some gestures (e.g., "stop," "move straight," "slow down") can achieve a high recognition rate of over 90%, whereas other gestures ("left turn," "right turn") have a low rate of less than 70%. The low recognition rates for these gestures are mainly due to the difficulty in accurately locating the arms in previous steps, as shown in Fig. 15.

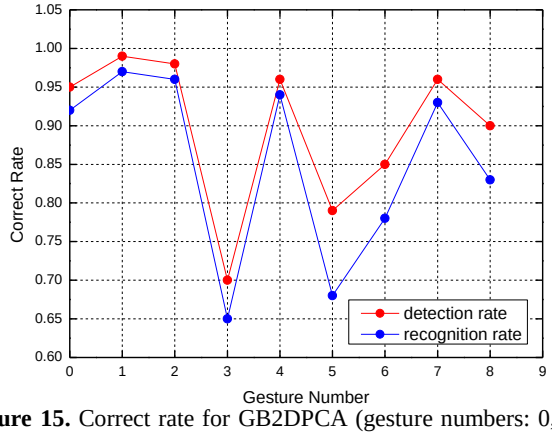



**Figure 15.** Correct rate for GB2DPCA (gesture numbers: 0, no gesture; 1, stop; 2, move straight; 3, left turn; 4, left turn waiting; 5, right turn; 6, change lane; 7, slow down; 8, pull over).

## 7. Discussions and conclusions

In this paper, we have proposed a very simple but effective algorithm for recognizing the gestures of traffic police in a complex scene. Key features of the proposed method are the use of a max-covering scheme to locate arms and the use of a Gabor feature-based 2DPCA to extract gesture features. Finally, the Gabor based 2DPCA is adopted to recognize police gestures. There are several advantages of the proposed method. First, the proposed method is based on a vision sensor, which is more convenient and cheaper than an on-body sensor-based method. Second, the method requires no special marks as is common in the motion-capture applications. Finally, the method integrates police features into the five-part body model searching to obtain a good arm location, and the police gestures can be recognized even if they are not performed perfectly.

However, the proposed algorithm also has some limitations: (1) The arm-located results of the proposed algorithm may not be correct while considering the side viewpoint of police. That is because the police torso is hard to accurately detect, which makes the five-part body model invalid in that case. (2) Although the proposed method can effectively exclude other person since only traffic police wear reflective vest, it can only deal with the situation with just one traffic police. For the image with more traffic police wearing reflective vests, our method shows a tendency to detect a wrong police torso. This disqualifies our algorithm from segmenting each traffic police in the same scene. (3) The police are required to be in focus, visible, and not blurry. If the vehicle is moving fast, the image may very well be blurry in that case, and our method will underestimate the foreground for the police. Nevertheless, we provide a new way to solve the problem of gesture recognition of Chinese traffic police only based on a vision sensor, which is rare in the literature. We intend to enhance the flexibility of the proposed algorithm in the future.

## Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 91220301), the China Postdoctoral Science Foundation (No. 2014M552154), the Hunan Planned Projects for Postdoctoral Research Funds (No. 2014RS4026), and the Postdoctoral Science Foundation of Central South University (No. 126648).

## References

[1] T. Yuan and B. Wang B, "Accelerometer-based Chinese traffic police gesture recognition system", Chin J Electron, 2010, 19(2): 270-274

[2] M. Singh, M. Mandal, and A. Basu, "Visual gesture

recognition for ground air traffic control using the Radon transform", In Proceeding of IEEE/RSJ IROS, Edmonton, Canada, 2005, pp 2586-2591

[3] H. Kang, C. W. Lee, and K. Jung, "Recognition-based gesture spotting in video games", Pattern Recogn Lett, 2004, 25(15):1701-1714

[4] L. W. Jin, D. D. Yang, and L. X. Zheng et al, "A novel vision-based finger-writing character recognition system", J Circuits Syst Comput, 2007, 16(3):421-436

[5] J. G. Liu, J. B. Luo, and M. Shan, "Recognizing realistic actions from videos 'in the wild'",

[6] Y. Song, D. Demirdjian, and R. Davis, "Tracking body and hands for gesture recognition: NATOPS aircraft handling signal database",

[7] B. Sapp, C. Jordan, and B. Taskar, "Adaptive pose prior for pictorial structures", In Proceeding of CVPR, San Francisco, USA, 2010, pp 422-429

[8] B. J. Zou, S. Chen, and C. Shi et al, "Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking", Pattern Recogn, 2009, 42:1559-1571

[9] S. Johnson and M. Everigham M, "Learning effective human pose estimation from inaccurate annotation", In Proceeding of CVPR, Colorado Springs, USA, 2011, pp 1465-1472

[10] Y. Zhu and K. Fujimura, "A Bayesian framework for human body pose tracking from depth image sequences", Sensors, 2010, 10:5280-5293

[11] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation", In Proceeding of CVPR, Colorado Springs, USA, 2011, pp 1465-1472

[12] M. Eichner and V. Ferrari, "Better Appearance Models for Pictorial Structures", In Proceeding of British Machine Vision Conference (BMVC), London, UK, 2009, pp 1-11

[13] Q. K. Le, C. H. Pham, and T. H. Le, "Road Traffic Control Gesture Recognition using Depth Images", IEEK Transactions on Smart Processing and Computing, 2012, 1(1): 1-7

[14] Z. Zhou, S. T Li, and B. Sun. "Extreme Learning Machine Based Hand Posture Recognition in Color-Depth Image", In proceedings of CCPR, 2014, pp. 1-10

[15] J. Tang, J. Luo., T. Tjahjadi, and Y. Gao, "2.5D Multi-View Gait Recognition Based on Point Cloud Registration", Sensors, 2014, 14: 6124-6143

[16] Y. M. Huang, G. B. Zhang, X. Li, and F. P. Da, "Improved emotion recognition with novel global utterance-level features", Appl Math Inf Sci, 2011, 5(2):147-153

[17] K. M. He, J. Sun, and X. O. Tang, "Single image haze removal using dark channel prior", In Proceeding of CVPR, Miami, FL, USA, 2009, pp 1956-1963

[18] A. Elgammal, R. Durauswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance", Proc IEEE, 2002, 90(7):1151-1163

[19] F. Guo, Z. X. Cai, and J. Tang J, "Chinese Traffic Police Gesture Recognition in Complex Scene", In Proceeding of FCST-11, Changsha, China, 2011, pp 1505-1511

[20] X. Pan and Q. Q. Ruan, "Palmprint recognition using Gabor featured-based (2D)2PCA", Neurocomputing, 2008, 71:3032-3036

[21] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004, 26(1): 131-137

[22] M. Hu, "Visual pattern recognition by moment invariants", IEEE Trans Inf Theory 1962, 8(2):179-187

[23] A. Khotanzad and H. H. Yaw, "Invariant image recognition by Zernike moments", IEEE Trans Pattern Anal Mach Intell, 1990, 12(5):489-497