Pattern Anal Applic (2015) 18:403–418
DOI 10.1007/s10044-014-0383-9

INDUSTRIAL AND COMMERCIAL APPLICATION

Max-covering scheme for gesture recognition of Chinese traffic police

Zixing Cai · Fan Guo

Received: 2 June 2013 / Accepted: 21 May 2014 / Published online: 3 June 2014
© Springer-Verlag London 2014

Abstract We present a method to recognize gestures made by Chinese traffic police in complex scenes based on a max-covering scheme for driver assistance systems and intelligent vehicles. Gesture recognition is made possible by upper-body-part detection with a five-part body model. 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 the proposed 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 rotation joint angle or Gabor feature-based two-dimensional principal component analysis is used to recognize the gestures made by Chinese traffic police. A comparative study is proposed with other human pose estimation methods, which demonstrates that better recognition results can be obtained using the proposed method on a number of video sequences.

Keywords Gesture recognition · Chinese traffic police · Dark channel prior · Max-covering · Five-part body model

1 Introduction

Gesture recognition of Chinese traffic police has important meanings for driver assistance systems and intelligent vehicles. However, this is a daunting task and is rare in the literature. This is mainly because it is hard to accurately detect traffic police in an unpredictable environment, which makes the problem both complex and limited.

As we will describe shortly in more detail, 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 choosing appropriate features for recognizing the traffic police's gestures. 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 rotation joint angles or Gabor feature-based two-dimensional principal component analysis (2DPCA).

The remainder of this paper describes our algorithm in more detail. We begin by reviewing existing works on gesture recognition. In Sect. 3, we explain the data flow diagram in our system. The detailed procedure to recognize the gestures of traffic police is described in Sects. 4, 5, and 6, whereas in Sect. 7, we illustrate our experimental results. At the end of this paper, we draw our conclusions about this study.

2 Previous work

Gesture recognition can be divided into roughly 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 Yuan et al. [1], on-body sensors were fixed on the back of

Z. Cai · F. Guo (✉)
School of Information Science and Engineering, Central South University, Changsha 410083, Hunan, China
e-mail: guofancsu@163.com

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 digital camera 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. [2] used the Radon transform to recognize hand gestures used by air marshals for steering aircraft on the runway. 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. [3] 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. Jin et al. [4] developed a video-based system for recognizing characters written with a finger. It allows one to enter characters into the computer program by using the movement of a fingertip.

On the other hand, gesture recognition and human body modeling can be closely related problems 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 [5, 6]. State-of-the-art pose estimation methods [7–9] typically represented the human body as a graphical model composed of ten major body parts corresponding to the head, torso, and upper and lower limbs. Meanwhile, human body models have also been used for body-part detection as in [10].

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. 1. The first step is to detect traffic police in a complex scene. The 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. 2).

In the last step, some typical gestures of traffic police are recognized in two ways. On the one hand, the rotation angles of the shoulder and elbow joints of each frame are used to match the defined rotation angles of standard gestures [11]. On the other hand, besides our previous work, 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 in this paper. A comparative study and a quantitative evaluation are proposed with other algorithms, which demonstrate that better quality results can be obtained by the proposed method.

4 Traffic police detection and foreground modeling

Locating the traffic police is the foundation of analyzing his/her gestures, which can be achieved through two steps: (1) police feature extraction and (2) police foreground modeling.

4.1 Police feature extraction

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 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:

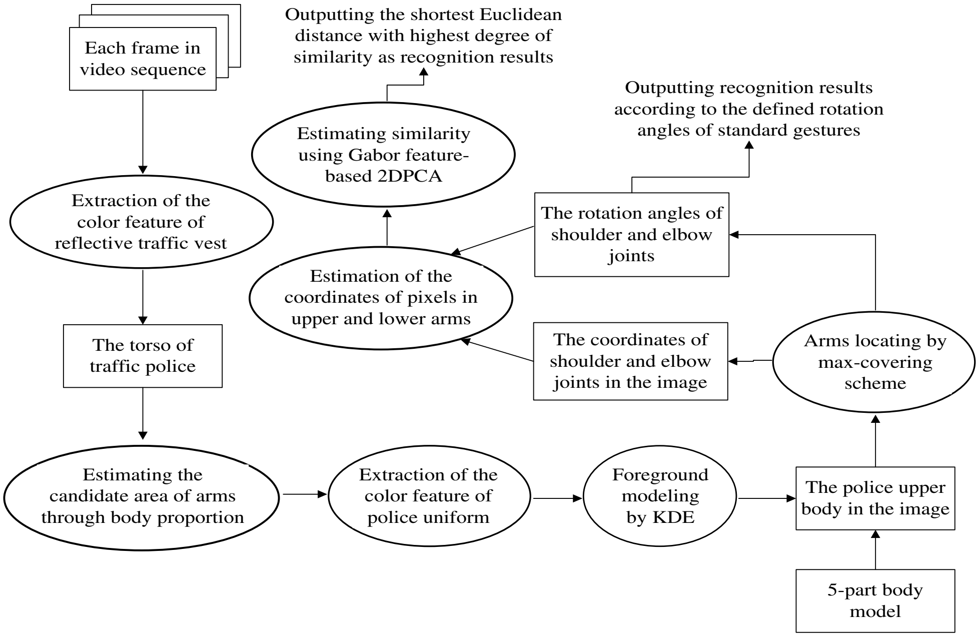

Each frame in video sequence
Extraction of the color feature of reflective traffic vest
The torso of traffic police
Estimating the candidate area of arms through body proportion
Extraction of the color feature of police uniform
Foreground modeling by KDE
The police upper body in the image
5-part body model
Arms locating by max-covering scheme
The coordinates of shoulder and elbow joints in the image
The rotation angles of shoulder and elbow joints
Estimation of the coordinates of pixels in upper and lower arms
Estimating similarity using Gabor feature-based 2DPCA
Outputting the shortest Euclidean distance with highest degree of similarity as recognition results
Outputting recognition results according to the defined rotation angles of standard gestures

Fig. 1 Diagram of data flow in our system

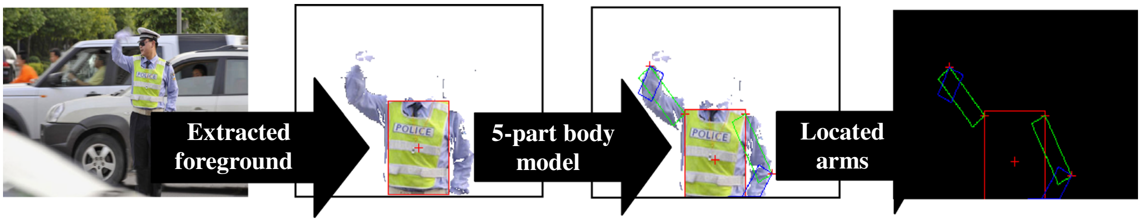


Fig. 2 Arm location by using a five-part body model

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

Thus, in Fig. 3a, for example, the vest, the tree, and the plant are green. Because of that, these objects were extracted by color thresholding. One example of thresholding can be observed in Fig. 3b. 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 [12], which was proposed to solve the de-hazing problem, is used here to further extract the vest. Formally, for an image J, its dark channel J^{dark} is defined by

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

where J^{c} is a color channel of J and $\Omega(x)$ is a local patch centered at x. The dark channel of Fig. 3b is shown in Fig. 3c. A detected pixel, x, will be considered to be part of the reflective vest only if $J^{\text{dark}}(x) > T$, as shown in Fig. 3d. The value of T is application based. We have decided to keep it fixed at 85 for all results reported in this paper. Figure 3e shows the result in the red bounding box. Note that if the height of the bounding box is too small (e.g., <1/20 of the image height), which means no police is detected, then there is no need to do the following steps.

4.2 Police foreground modeling

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

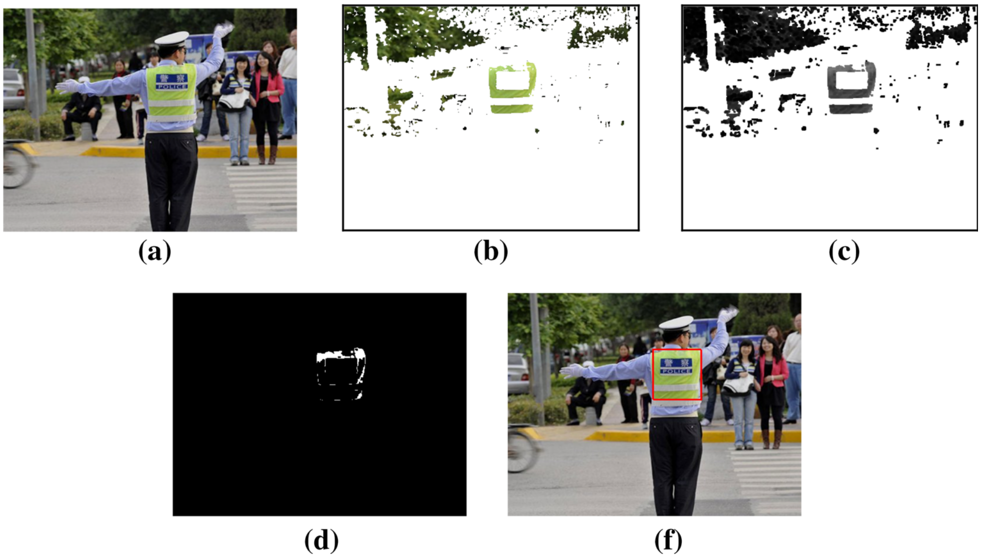


Fig. 3 Reflective vest extraction

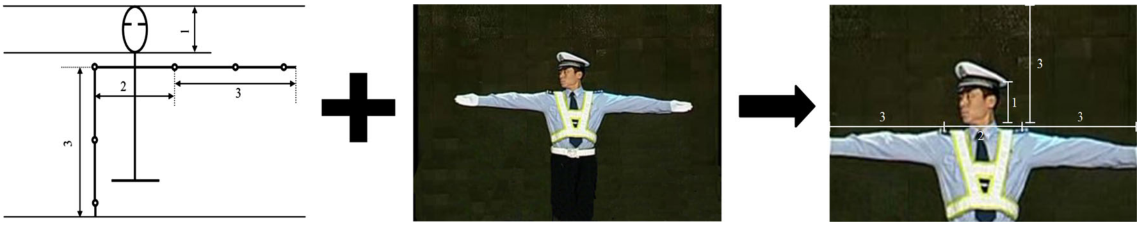


Fig. 4 The proportion of the human upper body

search area for the arms of traffic police. The proportion of the human upper body and the search area is depicted in Fig. 4.

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. Several segmentation methods could be used to separate the regions of the background from the regions of the police's arms and torso. However, we found in our experiments that it is better to estimate one distribution for the background and one distribution for the foreground using a kernel density estimator [13]. Assuming that the police's torso will be centered in the red bounding box (Fig. 5a), we first extract the pixels that satisfy the color threshold constraints: $b - g > 0.05$ and $b - r > 0.05$, as shown in Fig. 5b. 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. 5c, we use the top 20 pixels that are close to the center point of the box as the samples to estimate the probability distribution function (PDF) of the foreground. Let $x_1, x_2, \ldots, x_{20}$ be a sample of intensity values for a pixel. Given the intensity of target pixel x_t, we can estimate the density as

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

In Eq. 3, σ_j is a suitable bandwidth for R, G, and B three-color channel. In our experiment, σ_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. 5d.

5 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. In the following, we show how the location of the arms can be formulated as a max-covering problem. Here, we do not require the traffic police to wear special clothing or be instrumented with marks as is common in the pose estimation.

Fig. 5 Estimated police foreground image

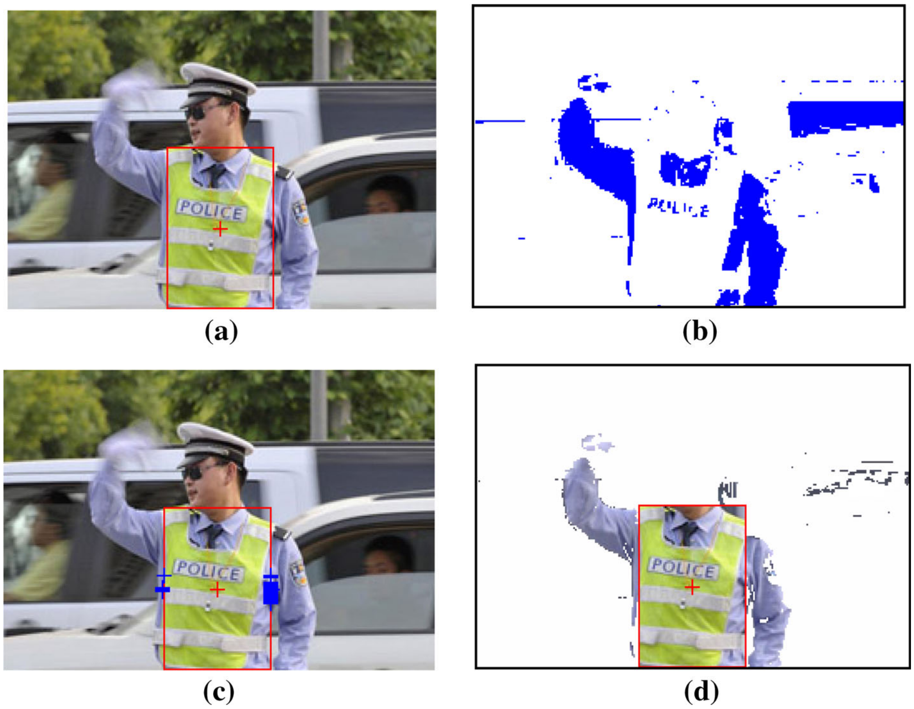
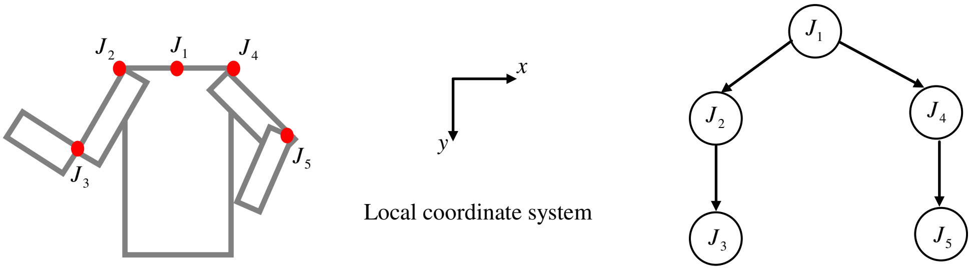


Fig. 6 Left: five-part body model; right: the tree structure of the five-part body model



5.1 Body model

We represent traffic police by using a five-part body, which is inspired by the widely used 10-part body model, which includes the head, the torso, the upper arms, the forearms, and the upper and the lower legs. 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. 6. 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. Figure 6 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. 6, 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.

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

5.2 Locating arms with the max-covering scheme

Each arm candidate covers some pixels in the foreground image. Intuitively, the arm tiles should cover foreground

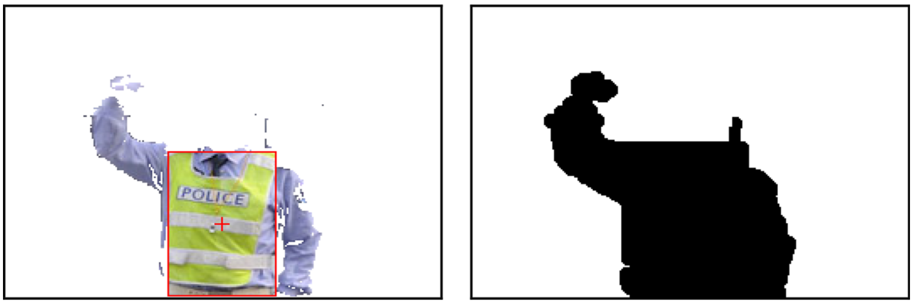


Fig. 7 Foreground image and its closed binary image

pixels as much as possible. Thus, locating the arm is performed with local research in the foreground image by using 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 Sects. 4.1 and 4.2, the max-covering scheme can be formulated as the following optimization problem:

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

where ϕ 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 θ, s, and r are used to control the coverage rate. The value of θ 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 θ 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 ϕ 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 ϕ 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 θ and s. Thus, a morphological operation is used here to tackle this problem. An illustrative example is shown in Fig. 7.

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. 8a, 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 θ 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. 8b, the image coordinate system of the pixel P_{R} that belongs to the right upper arm can be expressed as

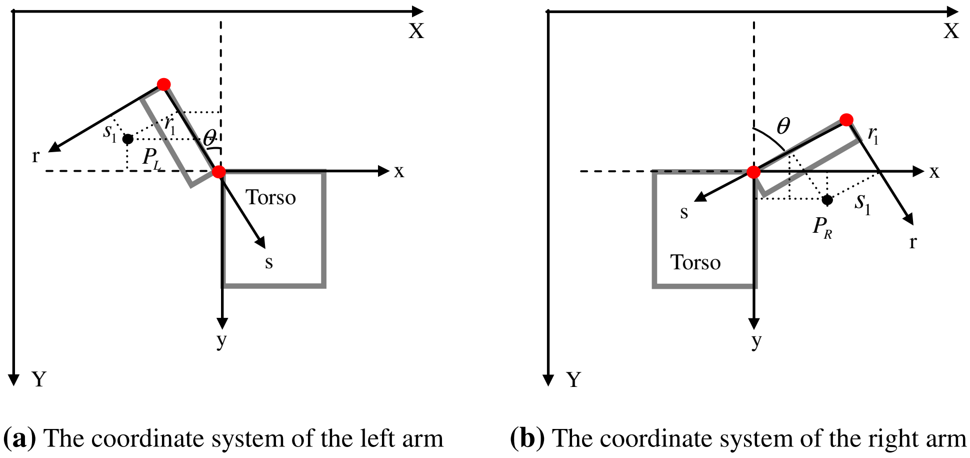
$$\begin{cases} u = x + (r_1 \cos\theta + (s - s_1)\sin\theta) \\ v = y + (r_1 \sin\theta - (s - s_1)\cos\theta) \end{cases}. \tag{7}$$

Therefore, the variable rectangles of upper arms can be obtained by adjusting different θ 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. 6, 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. Equations 6 and 7 are also used to obtain the variable rectangles of the forearms.

6 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:

Fig. 8 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

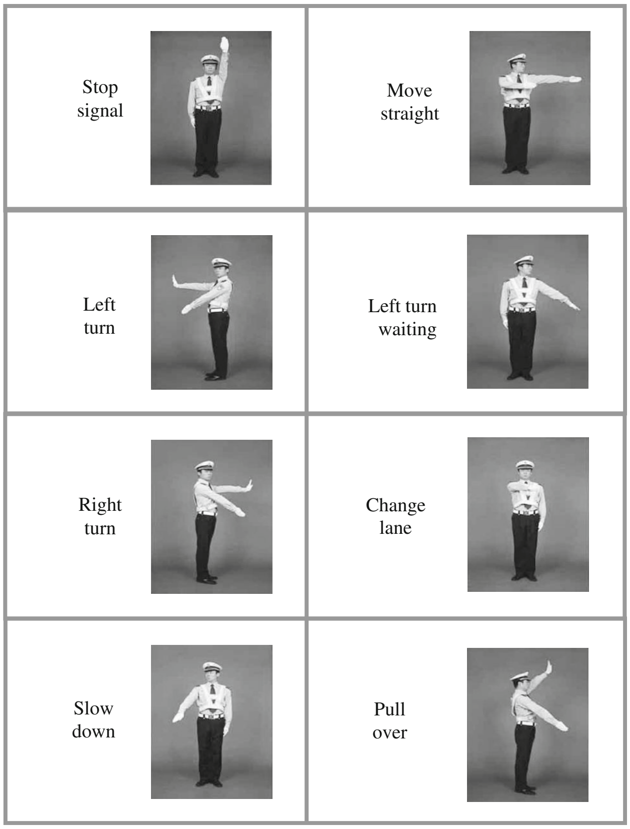


Fig. 9 Chinese traffic police gestures

(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. Figure 9 shows the eight gestures.

6.1 Recognizing gestures by using rotation joint angles

It can be seen from Fig. 9 that these gestures need the upper arms and forearms kept at certain angles and pointed in a vertical direction by rotating around shoulder or elbow joints, so the rotation joint angles are used to recognize gestures, which makes it easy to add a new gesture without changing the existing angles. Since the gestures may not be performed perfectly in a real situation, we set the angles in a certain range and not a fixed value. Let θ_i $(i = 1\ldots4)$ denote the rotation angle related to each arm for the gestures. Information about θ_i is provided in Table 2. Here, the rotation angles of two gestures, "stop" and "move straight (leftward or rightward)", are given in the table as examples. These two gestures are most important for intelligent vehicles.

6.2 Estimating similarity by using Gabor feature-based 2DPCA

Once the police's arms are located, a method of Gabor feature-based 2DPCA [14], which is used for palm print recognition and proved to be one of the best algorithms for object recognition, is adopted as another way to accomplish the gesture recognition task with appropriate parameters. Gabor filters can provide robust features against varying brightness and contrast of images. However, the procedure for feature coding and matching by pixels requires too much time and memory. Thus, statistical approaches, such as principal component analysis (PCA) or 2DPCA, can be used here to obtain useful features in gesture recognition. PCA is a useful statistical technique that has found application in fields such as pattern recognition and image compression and is a common technique for finding patterns in data of high dimension. The method is a powerful tool for analyzing data. PCA finds the collection of certain normalized orthogonal axes that indicate each direction of the maximum covariance for input data. However, the method is based on two assumptions: (1) the dimensionality of data can be efficiently reduced by linear transformation and (2) most information is contained in those directions where input data variance is maximum. As it is

Table 2 Rotation joint angle related to each arm

Gesture	Left upper arm (θ_1)	Left forearm (θ_2)	Right upper arm (θ_3)	Right forearm (θ_4)
Stop signal	$[0°,\ 10°]$	$[\theta_1 - 10°,\ \theta_1 + 10°]$	$[170°,\ 180°]$	$[\theta_3 - 10°,\ \theta_3 + 10°]$
Move straight (leftward)	$[80°,\ 110°]$	$[\theta_1 - 30°,\ \theta_1 + 30°]$	$[100°,\ 175°]$	$[\theta_3 + 30°,\ \theta_3 + 160°]$
Move straight (rightward)	$[100°,\ 175°]$	$[\theta_1 + 30°,\ \theta_1 + 160°]$	$[80°,\ 110°]$	$[\theta_3 - 30°,\ \theta_3 + 30°]$

evident, these conditions are by no means always met, and the directions maximizing variance do not always maximize information. A straightforward image projection technique, called 2DPCA, is developed for image feature extraction. As opposed to conventional PCA, 2DPCA is based on 2-D matrices rather than 1-D vectors; that is, the image matrix does not need to be previously transformed into a vector. Instead, an image covariance matrix can be constructed directly using the original image matrices. In contrast to the covariance matrix of PCA, the size of the image covariance matrix using 2DPCA is much smaller. As a result, 2DPCA has two important advantages over PCA. First, it is easier to evaluate the covariance matrix accurately. Second, less time is required to determine the corresponding eigenvectors. Therefore, 2DPCA is adopted in the proposed algorithm to extract features from given images.

Taking together both Gabor filters and 2DPCA, we use the Gabor feature-based 2DPCA for gesture recognition of Chinese traffic police. 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. The method is not only robust to illumination and rotation but also efficient in feature matching for the gestures of Chinese traffic police.

6.2.1 Gabor filter bank for feature extraction

2-D Gabor has the following general form:

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

where $i = \sqrt{-1}$, u is the frequency of the sinusoidal wave, and θ is the standard deviation of the Gaussian envelope. σ_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 [15], Zernike invariant moments [16], and others.

Table 3 Recognition rate of Gabor filter with different orientations and scales

Scale	Orientation				
	4	5	6	7	8
4	95.24	96.20	96.13	96.11	96.57
5	96.47	97.23	97.06	96.05	96.62
6	96.72	97.82	98.81	98.54	97.89
7	96.11	96.76	98.77	98.06	97.91
8	96.43	97.58	97.99	98.56	98.11

For Gabor filter, different orientations and scales can extract different image features and thus can obtain different recognition rates. Table 3 shows the recognition rate with different combinations of orientations and scales for 300 testing images. One can clearly see 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

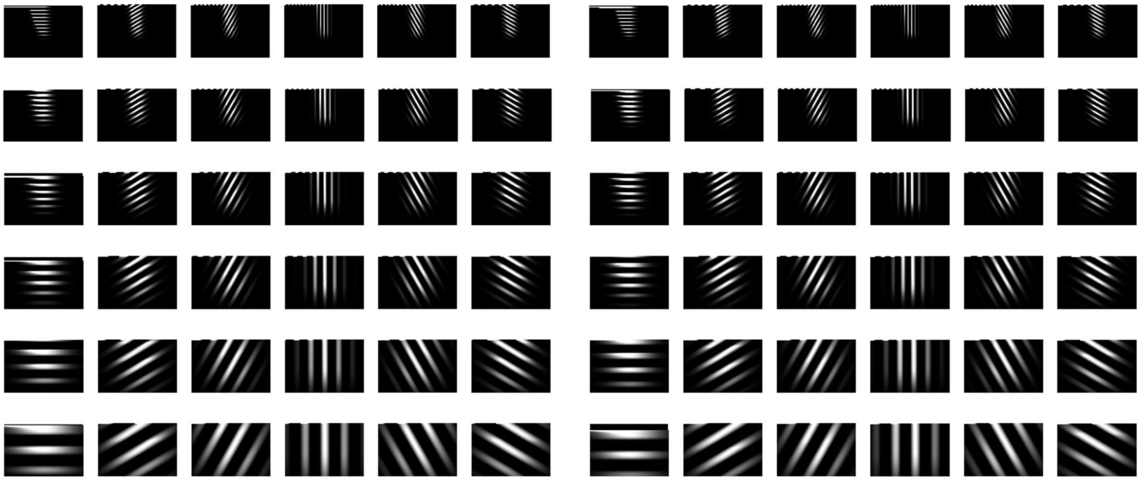
$$u = 0.2592/\sqrt{2}^{v} \qquad v = 0, 1, \ldots, 5. \qquad (9)$$

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

Thus, 36 Gabor filters are selected for feature extraction. Figure 10a, b shows the real and imaginary parts of the Gabor filter bank with six scales and six orientations.

Suppose that there are N 60 × 90 training police arm location images denoted by matrices as $A_i\,(i = 1, 2, \ldots, N)$. The convolution of the Gabor filter bank and image A_i yields Gabor feature matrices $H_i(v, k)$ $(v = 0, \ldots, 5;\ k = 1, \ldots, 6)$. Figure 11 shows the amplitude of the Gabor filter bank for the input arm-located image on the left.

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

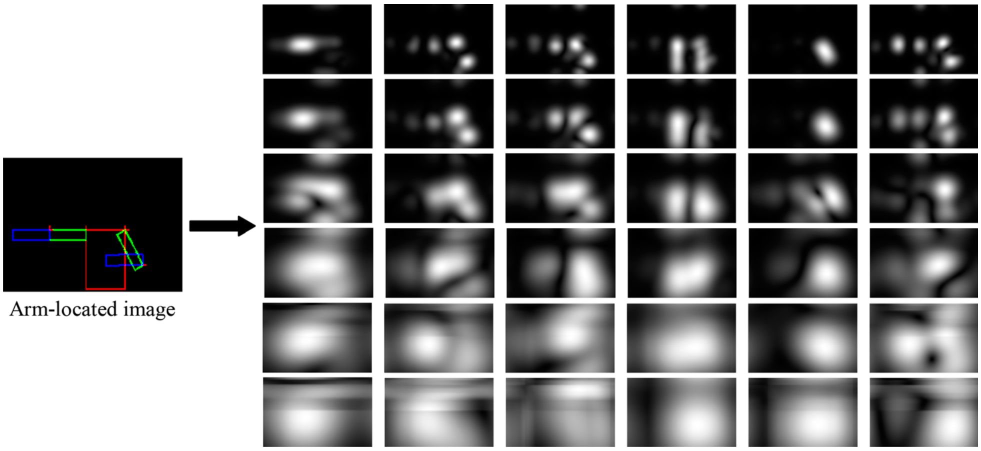


(a) Real part image

(b) Imaginary part image

Fig. 10 The real and imaginary parts of the Gabor filter bank with six scales and six orientations

Fig. 11 The amplitude image of the Gabor filter bank with six scales and six orientations



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

The Gabor feature space X is constructed by all the Gabor feature matrices of training samples in the row direction $X = \{X_1, X_2, \ldots, X_N\}$, 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.

6.2.2 2DPCA for dimensionality reduction of Gabor feature space

In 2DPCA, the covariance matrix G can be evaluated by

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

where

$$\bar{X} = \frac{1}{N}\sum_{i} X_i. \tag{13}$$

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

$$R = [r_1, \ldots, r_d]. \tag{14}$$

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 360×10, and the ultimate dimension of a Gabor feature vector is reduced from $36 \times 600 = 21{,}600$ to $360 \times 10 = 3{,}600$.

6.2.3 Gesture recognition using Gabor feature-based 2DPCA

After extracting features using 2DPCA, a nearest-neighbor classifier is adopted for classification. Supposed that traffic police gesture category c_i ($i = 1, 2,\ldots, p$) has N_i template training samples $B_j^{(i)} = (R_j^{(i)}, R_j^{(i)}, \ldots R_m^{(i)})$, $(j = 1, 2, \ldots, N_i)$ $N = \sum_{i=1}^{p} N_i$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

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

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 $D_{\mathrm{m}}(B) = \min\limits_{\substack{i\in(1,2,\ldots,p)\\ j\in(1,2,\ldots,N_i)}} \{D_i(B, B_j^{(i)})\}$ and $D_{\mathrm{m}}(B) \leq T$, then $B \in c_{\mathrm{m}}$, else B is not a traffic police gesture. T is called similarity threshold, and we set $T = 1.35$ in our experiment.

A simple illustrative example is shown in Figs. 12 and 13. The arm location results of standard police gestures constitute the template database. Figure 12 shows some sample templates for the gestures "stop", "move straight (left/right)", and "no sign". Four arm location results obtained by real captured photos are randomly chosen for testing, as shown in Fig. 13. From the Euclidean distance obtained by the above method, we can arrange the standard "move straight (left)" (see Fig. 12b) and four testing images in decreasing order of similarity: Fig. 13d, a, b, and c. Thus, we deduce that Fig. 13d indicates the "move straight (left)" gesture. This confirms our observation in real captured images.

7 Experiments and results

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 or located by a previous semiautomatic method. The assumption the proposed algorithm makes is that traffic police are seen approximately from a frontal viewpoint.

In the first experiment, we present results to show how the imprecise detection of a traffic police's torso affects the results of arm location and discuss which factors affect the precise detection of traffic police. The second experiment is performed to test the effectiveness of the two proposed gesture recognition methods. In this experiment, the location of the arms was automatically marked. In the last experiment, we present the comparative results and compare the proposed methods to the tree stick model method. In these experiments, the corresponding located arms are represented as rectangles in an automatic way or as a tree stick in a semiautomatic way.

7.1 How precision in the detection of the police's torso affects the result of arm location

It is important to note that imprecision in the detection of a traffic police's torso may greatly affect the results of arm location. In this section, some experiments for quantifying the influence of this detection on the final results are presented. Here, we use three different precise torso locations to analyze the effect of torso detection on the result of arm location. In the experiments, the test video sequences were captured by a Canon digital camera (S80) from a real driving scenario or from the Internet. Sample results of arm location using the proposed method are shown in Fig. 14. It shows the results of arm location under the three different

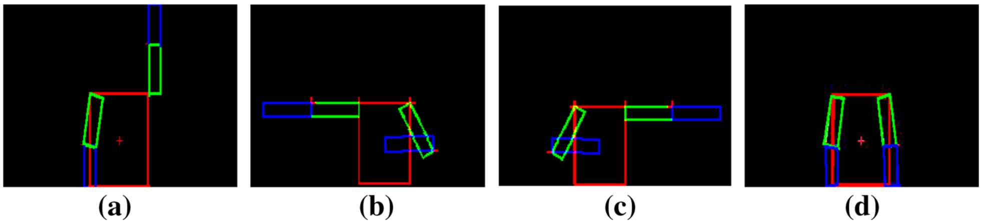


Fig. 12 Template examples: **a** arm location image of "stop", **b** arm location image of "move straight (*left*)", **c** arm location image of "move straight (*right*)", and **d** arm location image of "no sign"

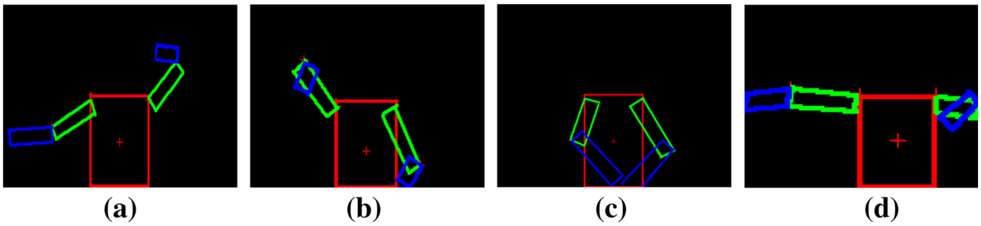


Fig. 13 Arm location examples with different Euclidean distances (D): **a** $D = 1.5931$, **b** $D = 1.7038$, **c** $D = 1.8143$, and **d** $D = 1.2933$

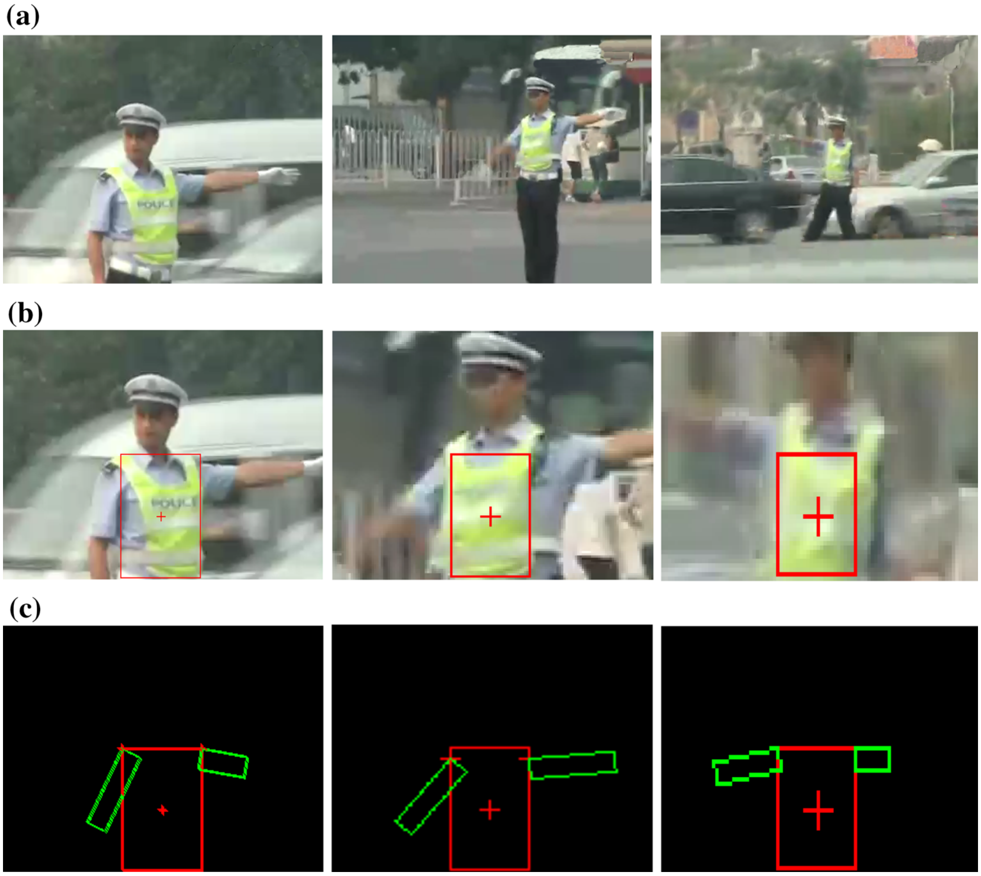


Fig. 14 Results of arm location for different distances between the police and the camera: **a** original images, **b** the results of detecting the police's torso, and **c** the results of the corresponding arm location

situations. Note that the forearm rectangles are not displayed since the foreground extraction is based on the color segmentation, and the traffic police wears a short-sleeved shirt. In Fig. 14, the images under the first column indicate the marked torso location with the highest precision. In the images under the second column, the torso location varies a little from the ground truth, and in the images under the third column, the error between the located torso and the corresponding ground truth is very great. We observe that the more precise the detection of the police's torso, the more precise the location of the police's arms. Thus, we can deduce that precision in regard to arm location depends on some factors, including camera resolution and the distance between the traffic police and the camera. The higher the resolution of the digital camera, the more precise the arm location will be, and the closer the police is to the digital camera, the more precise the arm location will be, as shown in Fig. 14.

7.2 Recognizing the gestures while the arm positions are located

We thoroughly tested our proposed method over several challenging video sequences captured by a digital camera or from the Internet. In the following, we evaluate the performance of two proposed algorithms with correct recognition rates 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, as shown in Fig. 15. For video gesture recognition, transitional gestures are not considered; only the final standard gestures are considered as recognition results, manual labeling or computer

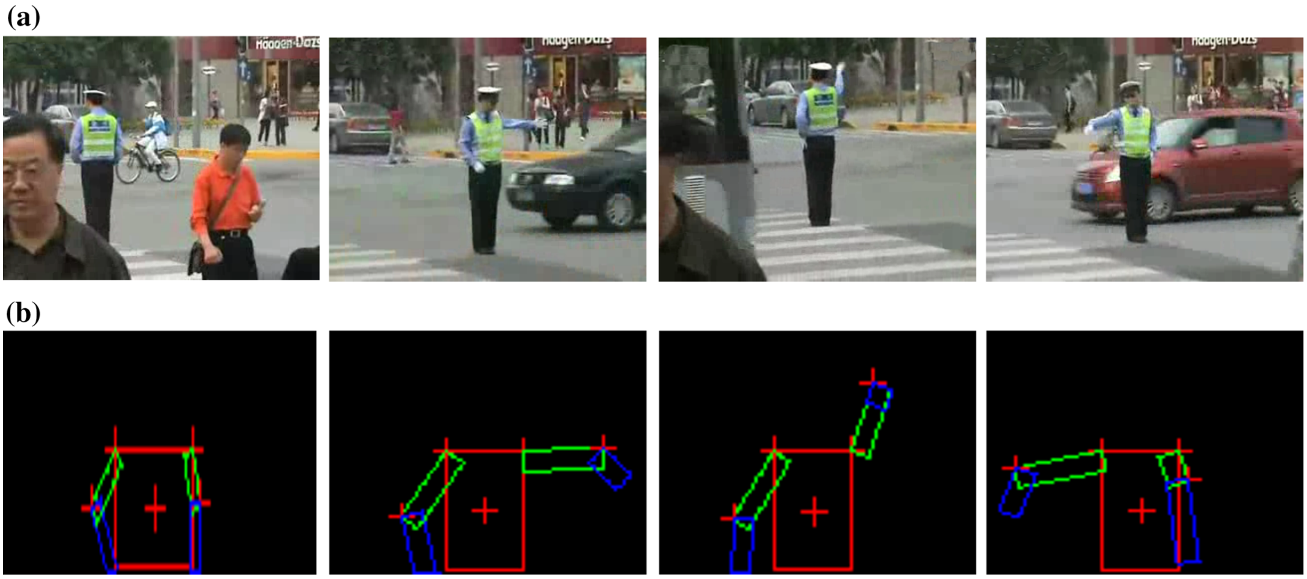


Fig. 15 Sample results of arm location using the proposed method: **a** original images, **b** arm location results

recognition process notwithstanding. The experiment's test environment is introduced as follows.

We perform the recognition algorithm by executing MATLAB on a PC with a 3.00 GHz Intel Pentium Dual-Core Processor. The frame resolution of the video is 320 × 240, and the typical torso of the traffic police appears as 35–42 pixels in height. The gestures in the video 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, as shown in Fig. 15. In the following, we respectively adopt the proposed two methods to recognize police gestures in the testing video. Note that 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, like in [17], 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.

7.2.1 *Recognition with rotation joint angles*

Recognizing police gestures with a rotation joint angle is very simple but effective. Figure 16 shows the recognition results of manual annotations and the rotation angle method.

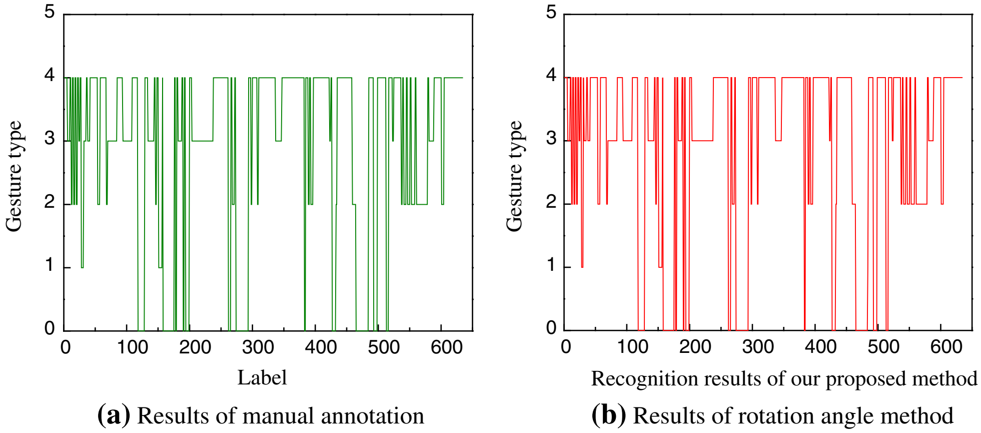
In the experiment, we defined the following errors in order to analyze the reasons for the wrong recognition results, as shown in Table 4:

- Omission error: mistook existing gesture for no gesture
- Substitution error: mistook one gesture for another.

It is clear from Table 4 that the rotation angle method achieved a high recognition rate even when the gestures are not very standard. The reason for the omission errors and substitution errors is that the algorithm selects the two vertices of the torso rectangle as the position of shoulder joints to locate the upper arms. Thus, when the algorithm only detects a small part of the vest as the police's torso, the rotation angle and the length of the arm will be wrong, and errors will be produced. For the rotation joint angle algorithm, the main advantage is its speed; its computational time for each frame of the video is less than one second in the MATLAB environment. However, only two gestures, "stop" and "move straight", are tested here. Since the rotation joint angles of other gestures only differ slightly among themselves, they are easily confused using this method. Thus, we propose a method using Gabor feature-based 2DPCA for recognizing gestures.

7.2.2 *Recognition with Gabor feature-based 2DPCA*

In this experiment, the correct recognition rate and testing time of the proposed Gabor feature-based 2DPCA (GB2DPCA) are investigated. The parameter setting of the Gabor filter is depicted in subsection 6.2.1, and the downsampling rate, ρ, remains 3 for the Gabor feature-based algorithm.



(a) Results of manual annotation **(b)** Results of rotation angle method

Fig. 16 Gesture recognition results by using rotation joint angles (gesture numbers: *0* no gesture, *1* stop, *2* left turn, *3* right turn, *4* other gestures)

Table 4 Correct recognition of police gestures

	Total number	Omission error	Substitution error	Correct recognition	Correct rate (%)
Proposed algorithm	634	12	19	603	95

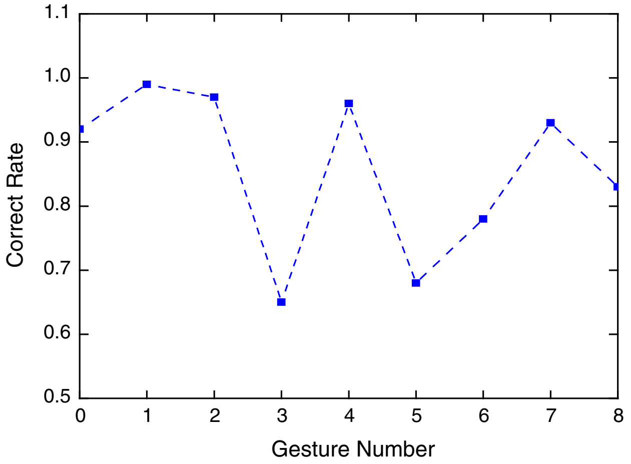


Fig. 17 Correct rate for GB2DPCA (gesture numbers: *0* no gesture, *1* stop, *2* move straight, *3* left turn, *4* left turn waiting, *5* right turn, *6* change lane, *7* slow down, *8* pull over)

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 above video sequences can be regarded as the testing image of gesture recognition. Figure 17 presented 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. 17. 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 <70 %. The low recognition rates for these gestures are mainly due to the difficulty in accurately locating the arms in previous steps. For GB2DPCA, the computational time is approximately 10 s per frame.

7.3 Comparison to the human pose estimation method based on tree stick model

Since the problem we are dealing with is considered quite novel, there are no directly similar methods to compare our results with. For instance, techniques in [1] are based on an accelerometer; however, equipping every member of the traffic police with the on-body sensor to command intelligent vehicles is impractical, and therefore, such a method does not fit in this environment. However, although the problem has not been the focus of the literature, substantial advances in human pose estimation have been made. In this section, we have shown that the proposed method to recognize gestures of Chinese traffic police can present encouraging recognition results. To prove the efficiency of the proposed method, we will compare it with the widely used tree stick model method [18]. The accuracy of the methods was compared using challenging test sequences.

For the tree stick model method [18], 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 Fig. 18b, c.

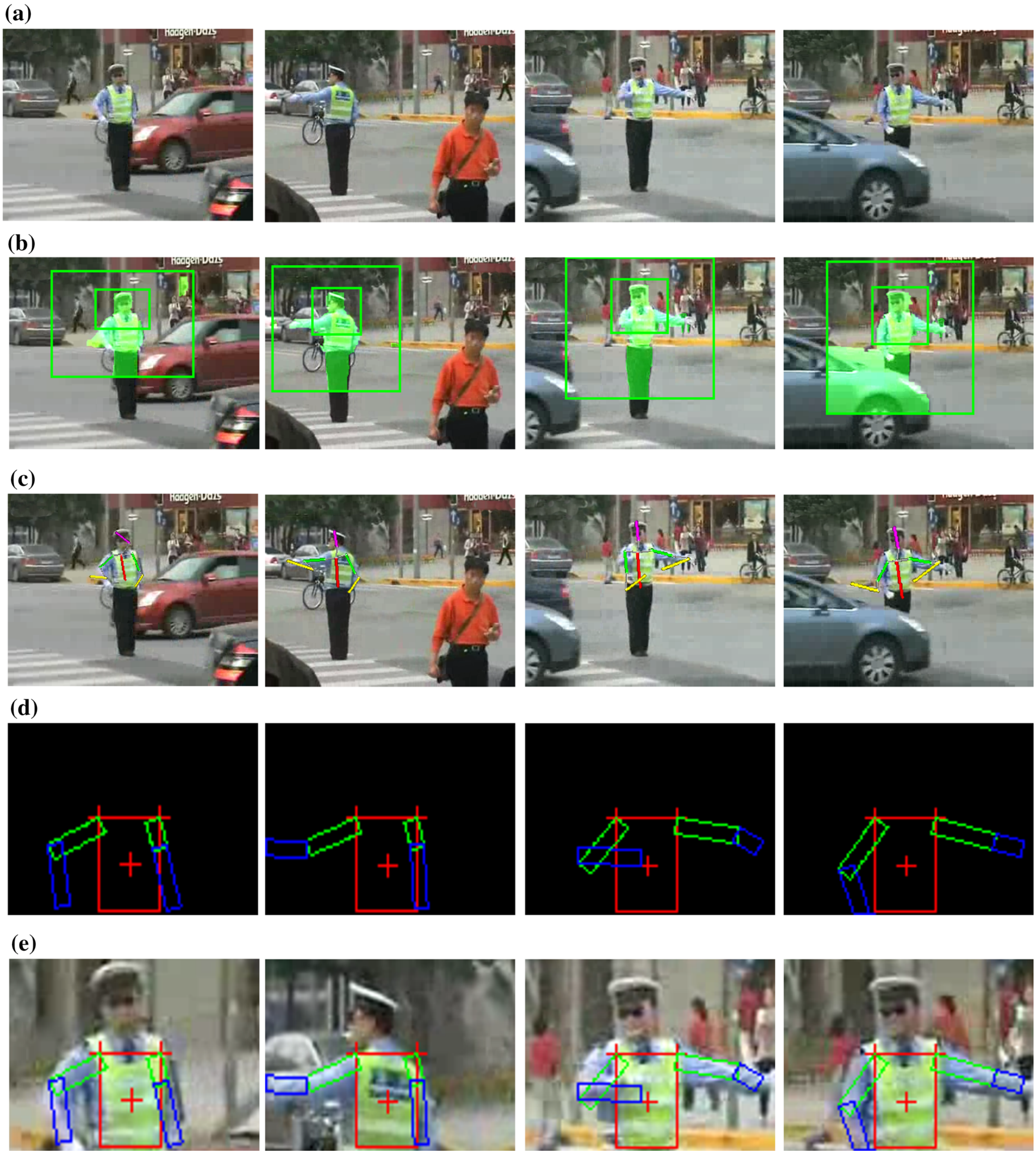


Fig. 18 Comparison between the tree stick model method and the proposed arm location method: The second and *third rows* show the upper-body poses for four frames obtained using the stick model method. The results for the proposed arm location method are shown in the last two rows. The images are as follows: **a** original frames, **b** the input of the tree stick model method, **c** body-part detection results by the tree stick model method, **d** and **e** body-part detection results by the proposed method

For our comparison with the tree stick model method, upper-body-part detection was performed on the original video sequences. As can be seen for the four selected frames in Fig. 18, our proposed method to recognize gestures was successful in locating the police's arms in the sequences. The arm box detector perfectly matched the

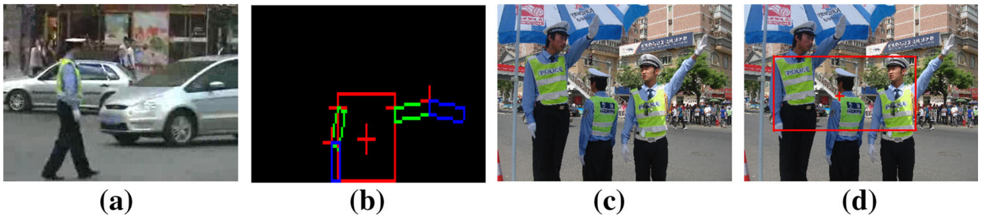


Fig. 19 Some failure examples

ground truth, as shown in Fig. 18e, whereas the tree stick model method was not able to locate the arms with the same efficiency (Fig. 18c). The reason for this is that the prior locations and the appearance transfer mechanism [17] 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 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 either the rotation joint angle method or the Gabor feature-based 2DPCA method. Furthermore, requiring user interaction also limits its use in police gesture recognition, as shown in Fig. 18b.

8 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 our proposed method are the use of a max-covering scheme to locate arms and the use of a rotation joint angle or Gabor feature-based 2DPCA to recognize 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 clothing or marks as is common in the motion-capture applications. Finally, the method exploits 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, as shown in Fig. 19b. 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 the road works since the color of their uniform is orange in most cases in China, 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, as shown in Fig. 19d. 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 algorithm in the future.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (No. 91220301, 61175064, 61273314), 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. Yuan T, Wang B (2010) Accelerometer-based Chinese traffic police gesture recognition system. Chin J Electron 19(2):270–274
2. Singh M, Mandal M, Basu A (2005) Visual gesture recognition for ground air traffic control using the Radon transform. In Proceeding of IEEE/RSJ IROS, Edmonton, Canada, pp 2586–2591
3. Kang H, Lee CW, Jung K (2004) Recognition-based gesture spotting in video games. Pattern Recogn Lett 25(15):1701–1714
4. Jin LW, Yang DD, Zheng LX et al (2007) A novel vision-based finger-writing character recognition system. J Circuits Syst Comput 16(3):421–436
5. Sapp B, Jordan C, Taskar B (2010) Adaptive pose prior for pictorial structures. In Proceeding of CVPR, San Francisco, USA, pp 422–429
6. Zou BJ, Chen S, Shi C et al (2009) Automatic reconstruction of 3D human motion pose from uncalibrated monocular video sequences based on markerless human motion tracking. Pattern Recogn 42:1559–1571
7. Johnson S, Everigham M (2011) Learning effective human pose estimation from inaccurate annotation. In Proceeding of CVPR, Colorado Springs, USA, pp 1465–1472
8. Zhu Y, Fujimura K (2010) A Bayesian framework for human body pose tracking from depth image sequences. Sensors 10:5280–5293
9. Johnson S, Everingham M (2011) Learning effective human pose estimation from inaccurate annotation. In Proceeding of CVPR, Colorado Springs, USA, pp 1465–1472

10. Lee MW, Nevatia R (2007) Body part detection for human pose estimation and tracking. In Proceeding of IEEE workshop on motion and video computing (WMVC), Texas, USA, pp 1–8
11. Guo F, Cai ZX, Tang J (2011) Chinese Traffic Police Gesture Recognition in Complex Scene. In Proceeding of FCST-11, Changsha, China, pp 1505–1511
12. He KM, Sun J, Tang XO (2009) Single image haze removal using dark channel prior. In Proceeding of CVPR, Miami, FL, USA, pp 1956–1963
13. Elgammal A, Durauswami R, Harwood D, Davis LS (2002) Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proc IEEE 90(7):1151–1163
14. Pan X, Ruan QQ (2008) Palmprint recognition using Gabor featured-based $(2D)^2$PCA. Neurocomputing 71:3032–3036
15. Hu M (1962) Visual pattern recognition by moment invariants. IEEE Trans Inf Theory 8(2):179–187
16. Khotanzad A, Yaw HH (1990) Invariant image recognition by Zernike moments. IEEE Trans Pattern Anal Mach Intell 12(5):489–497
17. Huang YM, Zhang GB, Li X, Da FP (2011) Improved emotion recognition with novel global utterance-level features. Appl Math Inf Sci 5(2):147–153
18. Eichner M, Ferrari V (2009) Better Appearance Models for Pictorial Structures. In Proceeding of British Machine Vision Conference (BMVC), London, UK, pp 1–11