Multimed Tools Appl (2017) 76:8915–8936
DOI 10.1007/s11042-016-3497-9

Gesture recognition of traffic police based on static and dynamic descriptor fusion

Fan Guo[1] · **Jin Tang**[1] · **Xile Wang**[1]

Received: 17 August 2015 / Revised: 19 March 2016 / Accepted: 24 March 2016 /
Published online: 15 April 2016

Abstract We present a method to recognize gestures made by Chinese traffic police based on the static and dynamic descriptor fusion for driver assistance systems and intelligent vehicles. Gesture recognition is made possible by combining the extracted static and dynamic features. First, the point cloud data of human upper body in each frame of input video is obtained to estimate the static descriptor with 2.5D gesture model. Then, the dynamic descriptor is estimated by computing the motion history image of the input RGB video sequence. Finally, the above two descriptors are fused and the mean structural similarity index is used to recognize the gestures made by Chinese traffic police. A comparative study and qualitative evaluation are proposed with other gesture recognition methods, which demonstrate that better recognition results can be obtained using the proposed method on a number of video sequences.

Keywords Chinese traffic police · Gesture recognition · 2.5D gesture model · Motion history image · Descriptor fusion

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 recognize gestures when the arms of traffic police are perpendicular to the image plane, which makes the problem both complex and limited. For gesture recognition of Chinese traffic police, the task generally faces two challenges. One is choosing appropriate features for representing the traffic police's gestures. The problem is very hard because of the specific

✉ Jin Tang
tjin@csu.edu.cn

Fan Guo
guofancsu@163.com

[1] School of Information Science and Engineering, Central South University, Changsha, Hunan 410083, China

position of the police's arms. The other challenge involves adopting proper classification procedure to correctly recognize the traffic police's gestures.

Because of the importance of the gesture recognition algorithm, much work has been done. These methods 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 [21], 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 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. Nowadays, the vision sensor–based method has achieved both scientific and economic success. For example, Singh et al. [15] used the Radon transform to recognize hand gestures used by air marshals for steering aircraft on the runway. For the method, a binary skeleton representation of the human subject is computed. The Radon transform is used to generate maxima corresponding to specific orientations of the skeletal representation. Kang et al. [11] 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, and a novel gesture spotting method that combines gesture spotting with gesture recognition is proposed in the method. Liu et al. [13] presented a systematic framework for recognizing realistic actions from videos "in the wild". Both motion and static features are extracted from testing videos, and AdaBoost is chosen to integrate all the heterogeneous yet complementary features for recognition. Song et al. [17] proposed a unified framework for body and hand tracking, and a multi-signal gesture database is also introduced in their work. The framework presented in their work uses a stereo camera to collect 3D images, and tracks body and hand together. Guo et al. presented a method to recognize gestures made by Chinese traffic police using a max-covering scheme [2, 7, 8]. The scheme is proposed based on a key observation that body-part tiles maximally cover the foreground region and satisfy a body plan. Then, the rotation joint angle [2, 7] or Gabor feature based two-dimensional principal component analysis [2, 8] is used to recognize the police gestures. For the methods [2, 7, 8], only RGB camera is used to capture testing images and videos, and the 5-part body model constructed in these methods is a two-dimensional (2D) model. Therefore, these methods are hard to deal with the situation when the arms of traffic police are perpendicular to the image plane. The CALVIN research group has made great progress in human pose estimation [4–6]. For example, Ferrari et al. [6] proposed an approach that reduces the search space for human body parts and integrates spatial-temporal model covering multiple frames to refine pose estimates from individual frames with inference using belief propagation. Eichner et al. [4] presented a human pose co-Estimation method for joint pose estimation over multiple persons in a common but unknown pose. Eichner et al. [5] proposed a method for estimating the spatial layout of humans in still images—the position of the head, torso and arms. For the method, once a person is localized using an upper body detector, the search for their body parts can be simplified using weak constraints on position and appearance arising from that detection.

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. Many researchers 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 [14, 25]. For example, the adaptive pose priors proposed by Sapp et al. [14] is a semi-parametric approach that combines the tractability of pictorial structure inference with the flexibility of non-parametric methods. Zou et al. [25] presented a method to reconstruct human motion pose from uncalibrated monocular video sequences based on the morphing appearance model matching. The human pose estimation is made by integrated human joint tracking with pose reconstruction in depth-first order. State-of-the-art pose estimation methods [10, 24] 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. Johnson et al. [10] utilized Amazon Mechanical Turk and a latent annotation update scheme to achieve high quality annotations at low cost. Zhu et al. [24] presented a flexible Bayesian framework for integrating pose estimation results obtained by the methods based on key-points and local optimization. Eichner et al. [3] propose a novel body part appearance models for pictorial structures. They learned latent relationships between the appearances of different body parts from annotated images, which then help in estimating better appearance models on novel images. Besides, a 3D measuring device–Kinect has been used to construct human pose model for gesture or gait recognition as in [12, 19, 23]. For example, Le et al. [12] use a built-in depth sensor of Microsoft Kinect to capture the control gestures of traffic police officers in the form of depth images. A human skeleton is then constructed using a kinematic model, and the feature vector describing a traffic control gesture is built from the relative angles found amongst the joints of the constructed human skeleton. Zhou et al. [23] collect color-depth images with the Microsoft Kinect sensor to recognize hand posture. To accurately locate hands in the images with complex background, a depth histogram based adaptive thresholding method is adopted for the depth image and the Bayesian skin-color detection is performed for the corresponding color image. Then the two processed results are fused and refined with a region-growing method. Tang et al. [19] use a single Kinect camera to obtain point cloud data of a human body and construct 2.5-dimensional (2.5D) voxel gait model. Based on the model, the gait features can be extracted for identifying the human subject. In this paper, we also use Kinect camera to obtain the point cloud data of the police's upper body, and then construct a 2.5D gesture model for the gesture recognition of Chinese traffic police.

In this work, we obtain the static descriptor using the 2.5D surface model of the police's upper body constructed by a single Kinect camera, and exploit dynamic features of police gestures by introducing the motion history image (MHI). Thus, the final descriptor which we used to represent the traffic police's gesture features can be computed by fusing the static and the dynamic descriptors. With the fused descriptor, the police gestures are recognized by using the mean structural similarity (MSSIM) index.

The remainder of this paper describes our algorithm in more detail. In Section 2, we explain the data flow diagram in our system. The detailed procedure to recognize the gestures of traffic police is described in Sections 3, and 4, whereas in Section 5, we illustrate our experimental results. In Section 6, we discuss some critical issues related to the proposed method. At the end of this paper, we draw our conclusions about this study.

2 System overview

The basic idea of our algorithm is to recognize police gestures from the corresponding fused descriptor obtained by static and dynamic feature extraction. The static descriptor of each

frame of the video is estimated using a 2.5D gesture model, and the dynamic descriptor of the input video sequence is obtained with the computed MHI image. To obtain the 2.5D gesture model and the MHI image, a Microsoft Kinect sensor is adopted for this purpose.

Figure 1 depicts the algorithm framework of our gesture recognition system. As can be seen in the figure, the proposed algorithm is divided into three major steps. The first step is to extract the static feature by using a 2.5D gesture model. The step includes two sub-steps: (1) obtaining the point cloud data of human upper body in each frame image of the video and (2) estimating the static descriptor by doing the re-sampling with these point cloud data.

In the second step, the dynamic feature is extracted by computing the motion history image (MHI) of the input RGB video sequence. Since the kinect camera used for collecting data can provide not only the depth information, but also the RGB image, so the MHI is computed from the RGB images of the input video as the dynamic descriptor.

In the last step, the static descriptor and the dynamic descriptor are first fused, and then the typical gestures of traffic police are recognized in a simple but effective way. In this paper, we propose to use mean structural similarity (MSSIM) index by calculating the largest MSSIM value with the highest degree of similarity from the result of the descriptor fusion to recognize police gestures. A comparative study and a quantitative evaluation are proposed with other methods, which demonstrate that better quality results can be obtained by the proposed method.

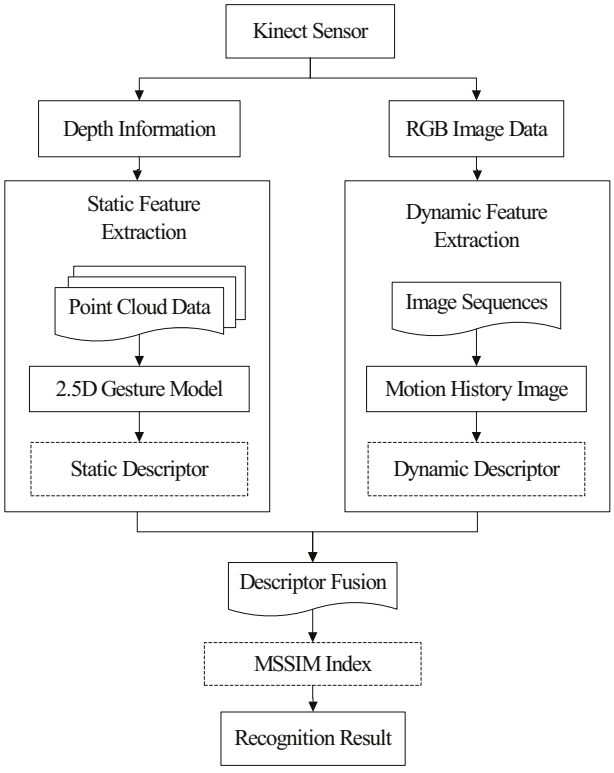


Fig. 1 Our gesture recognition framework

3 Feature extraction

Extracting the gesture feature is the foundation of analyzing his/her gestures, which can be achieved through three steps: (1) static descriptor estimation, (2) dynamic descriptor estimation and (3) static and dynamic descriptor fusion.

3.1 Static descriptor estimation

3.1.1 Construction of 2.5D gesture model

For static descriptor estimation, constructing a proper gesture surface voxel model is very important for representing the static feature. Here, 2.5D data that contains depth information is used to construct the voxel model, and a Kinect is used to capture the 2.5D data which is a simplified 3D (x, y, z) surface representation (Fig. 2). 2.5D data contains at most one depth value $d(x, y)$ which denotes the distance between the RGB image pixel (x, y) of a point on the upper body surface and the Kinect. 2.5D is a suitable trade-off solution between 2D and 3D approaches. It is restricted to a given viewpoint that is called 2.5D information [18].

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. The 2.5D data of the depth image and RGB image are used to construct a 3D voxel model by calculating all the 3D points from the measurement (x, y, d) in the depth image. 3D point cloud data are calculated using the Kinect geometrical model [16], i.e.:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{1}{c_1 d + c_0} dis^{-1} \left(K^{-1} \begin{bmatrix} x + u_0 \\ y + v_0 \\ 1 \end{bmatrix}, k \right) \tag{1}$$

where d is depth value along the z-axis, c_1 and c_0 are parameters of the model, u_0 and v_0 are respectively the shifted parameters of IR and depth images, dis is distortion function, k is distortion parameter of the Kinect IR camera and K is the IR camera calibration matrix.

Specifically, gesture silhouettes are first extracted from the depth image, and RGB images are then used to calculate all the 3D point cloud data for the gesture using Eq. (1). The 3D

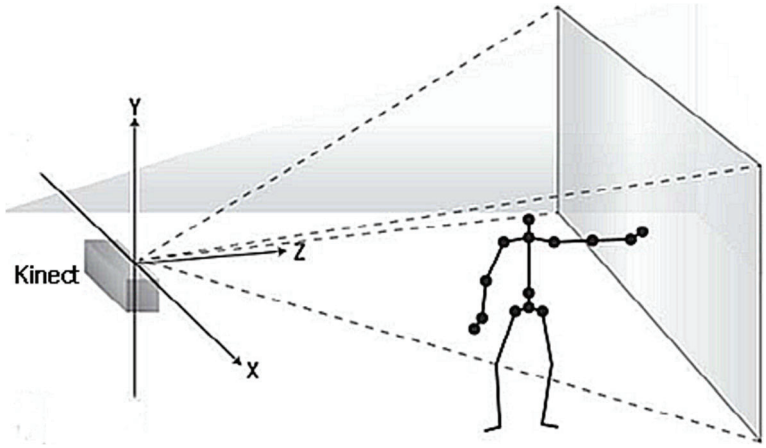


Fig. 2 World coordinates of the Kinect sensor-based system

point cloud gesture model is constructed for a given viewpoint by normalizing all the gesture point cloud data to 3D space. Since only a single Kinect depth camera is used, the gesture point cloud data includes only one side surface portion of the human upper body as shown in Fig. 3. We call it a 2.5D gesture model.

3.1.2 Extraction of gesture static feature

After constructing the 2.5D gesture model, the static gesture feature can be extracted by using the model. Specifically, the normalized point cloud data is firstly projected onto the XY plane into $N_I \times N_J$ blocks as shown in Fig. 4, where dx and dy are respectively the horizontal and vertical sampling intervals. Each point cloud data is located in the corresponding block (I, J), and each block may have several point cloud data. Since only a single Kinect camera is used, the point cloud data include only one side surface portion of the human upper body, and the surface point set is denoted by $S=[x, y, z]$, $(x,y) \in D$, where D is the projected grid from 2.5D gait surface onto the XY plane. In order to obtain the static feature by doing re-sampling operation, the mean value of z are calculated for all point cloud data located in the block (I, J), denoted by $z_{mean(I,J)}$, and the discrete surface point set is then obtained $S_{dis}=[I, J, z_{mean(I,J)}]$. Note that the construction of static descriptor is actually a procedure of projecting 3D point cloud to 2D image. The 2D image that represents the re-sampled point cloud is called gesture feature image, which is denoted by $G_f(x, y)$.

Besides, a complete police gesture includes several frames, and the static descriptor of each gesture is reconstructed by data from corresponding frames. Most often, the number of frames for each gesture is different, e.g., the sample left turn gesture has fourteen key frames while the stop gesture has only six key frames. It is thus difficult to automatically separate a complete gesture cycle and directly compare two gestures with different number of frames. In order to overcome these problems, as a preliminary process, we just manually separate each gesture cycle and extract key frames. Meanwhile, the final static descriptor $D_s(x, y)$ is computed as the average value of feature images of the key frames in each gesture cycle, as shown in Fig. 5. The process can be written as:

$$D_s(x,y) = \frac{1}{N}\sum_{i=1}^{N}\left(G_f(x,y)\right) \tag{2}$$

In (2), $G_f(x, y)$ is the gesture feature images of the key frames at pixel (x, y). N is the number of the key frames in a gesture cycle, and $D_s(x, y)$ is the final static descriptor.

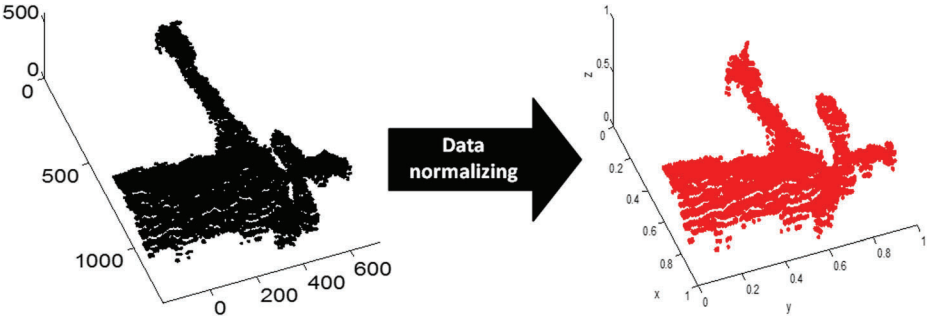


Fig. 3 The normalized point cloud data of human upper body

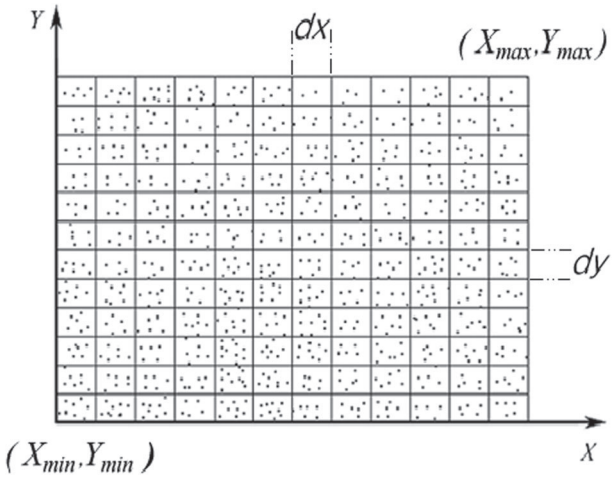


Fig. 4 Projection of normalized point cloud data onto blocks

3.2 Dynamic descriptor estimation

Ideally, depth maps will have highly accurate 3D information, in which case examining static descriptor would suffice to recognize gesture successfully. However, current depth sensor technology is limited in resolution (i.e., depth accuracy decreases exponentially as the distance gets further), so relying solely on the static 3D point cloud returned from the sensor will lead to an unsatisfactory result. Instead we also want to exploit dynamic descriptor of arm motion, and we do this by introducing the motion history image (MHI), in which each pixel value is a function of the recent motion in that location in the image. This often gives us useful information about dynamics of gesture, indicating how the gesture has occurred.

Intensity of each pixel in the MHI is a function of motion density at that location. One of the advantages of the MHI representation is that a range of times may be encoded in a single frame, and in this way, the MHI spans the time scale of human gestures [1]. Specifically, the MHI $H_\tau(x, y, t)$ can be computed from an update function $\psi(x, y, t)$:

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \psi(x,y,t) = 1 \\ \max(0, H_\tau(x,y,t-1)-\delta) & \textit{otherwise} \end{cases} \tag{3}$$

Here, (x, y) and t show the position and time, $\psi(x, y, t)$ signals object's presence (or motion) in the current video image, the duration τ decides the temporal extent of the movement (e.g., in terms of frames), and δ is the decay parameter. The result of this computation is a scalar-valued image where more recently moving pixels are brighter and vice-versa. In our experiment, the MHI is generated from a binarized image, obtained from frame subtraction, using a threshold ξ :

$$\psi(x,y,t) = \begin{cases} 1 & \text{if } D(x,y,t) \geq \xi \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $D(x, y, t)$ is defined with difference distance Δ as:

$$D(x,y,t) = |I(x,y,t) - I(x,y,t \pm \Delta)| \tag{5}$$

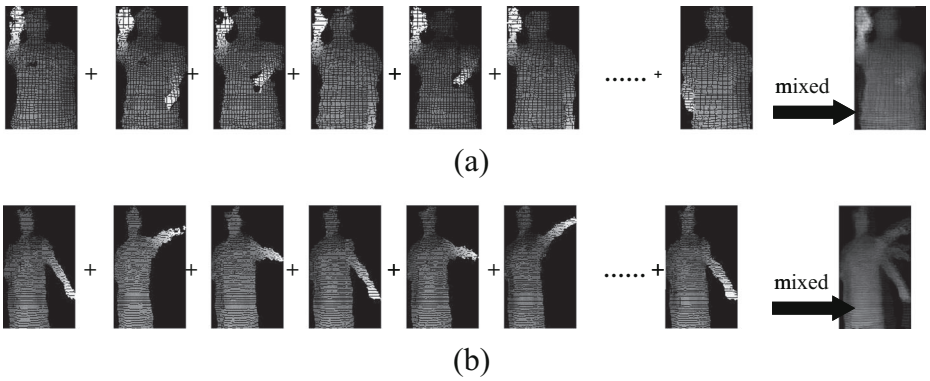


Fig. 5 The static descriptor obtained by mixing gesture feature images in a cycle

Here, $I(x, y, t)$ is the intensity value of pixel location with coordinate (x, y) at the tth frame of the image sequence. We can get the final MHI template as $H_\tau(x, y, \tau)$. Figure 6 presents the estimated MHI images for different police gestures. These illustrate clearly that a final MHI image records the temporal history of motion in it. Thus, the MHI image can be used as the dynamic descriptor $D_d(x, y)$ of input video sequence.

3.3 Descriptor fusion

The fusion between static descriptor and dynamic descriptor is accomplished through the weighted linear method. The fused descriptor $D_f(x, y)$ is thus computed as:

$$D_f(x,y) = w_1 \cdot D_s(x,y) + w_2 \cdot D_d(x,y) \tag{6}$$

where $D_s(x, y)$ and $D_d(x, y)$ are the static descriptor and dynamic descriptor, respectively. $D_f(x, y)$ is the result of descriptor fusion used for gesture recognition. w_1 and w_2 are the two weight values used to control the proportion of static descriptor and dynamic descriptor in the final descriptor fusion result, and $w_2 = 1\text{-}w_1$. To analyze and determine the value of the two weights, we first obtain the descriptor fusion results with different weight values for the eight standard gestures in the sample image database. Here, the parameter w_1 is set over the range [0.1, 0.9] by a certain interval, which wet as 0.1. Then, the descriptor fusion results for different gesture

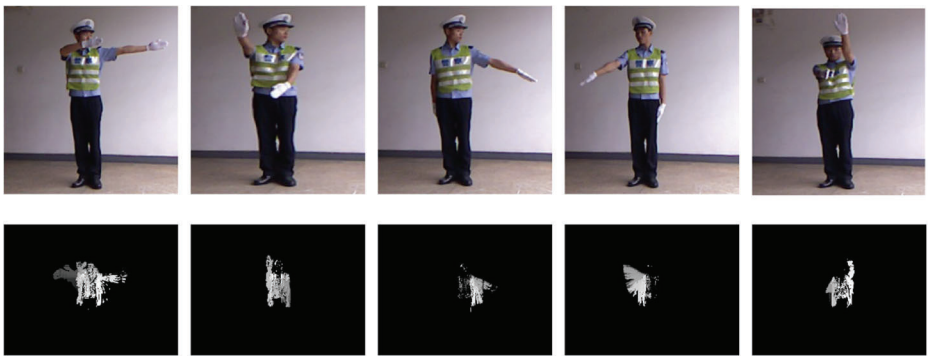


Fig. 6 The estimated MHI images for different police gestures

testing images are also estimated using the same parameter settings. An illustrative example is shown in Fig. 7. As can be seen in Fig. 7, as the value of w_1 increasing, the static descriptor of the gesture "move straight" becomes dominant in the final descriptor fusion results, while the dynamic descriptor of the gesture becomes more and more weak. When $w_1 = w_2 = 0.5$, both the static features and the dynamic features of police gestures can be well characterized in the final descriptor fusion results.

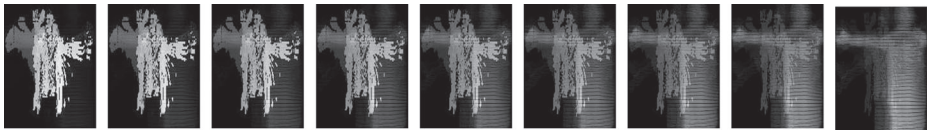
In the experiment, we find that as long as the proportion of the static descriptor and the dynamic descriptor keep the same for the sample image and the test image, meaning both the sample and test image have the same w_1 and w_2 value, the correct gesture categories can be surely determined for the test image. Table 1 shows the MSSIM index values obtained using different weight parameters for the gesture "move straight" (#2), in which "S" stands for the sample image. It can be seen from the table that the correct gesture category "2" is obtained through the maximum values of MSSIM index for all the combinations of w_1 and w_2.

Figure 8 also presents the MSSIM index value for the gesture "move straight" (#2) to better reveal the experimental results. One can clearly see that the curve corresponding to S2 lies in the top of the figure, while other curves seem near the bottom and close to each other, which indicates that the MSSIM index value for the testing gesture has relatively large differentiation and the testing gesture belongs to the gesture category "2". Therefore, from the experimental results, we can deduce that the MSSIM index corresponding to the correct gesture is already distinctive compared to others, regardless of the combinations of w_1 and w_2. The experimental results obtained with other gestures also confirmed the above conclusions.

Based on the above conclusions, the two weight parameters w_1 and w_2 are fixed to 0.5 for the sample image database and all the test images in our experiment. Figure 9 shows an example of descriptor fusion result. The static descriptor for "left turn waiting" gesture is shown in Fig. 9a and the corresponding dynamic descriptor is shown in Fig. 9b. Note that the arm is on the left in Fig. 9a, while in Fig. 9b the arm is on the right. Since the same process is performed on the sample and test image dataset, thus the image feature can be regarded as a reliable descriptor for gesture recognition even the arm is on different sides for the static and dynamic descriptors. Figure 9c is our final descriptor fusion result. One can clearly see that the static descriptor captures discrepancies in static pose, while the dynamic descriptor captures discrepancies in the dynamics of motion. The fused descriptor combines both descriptors by setting the weight of each term empirically. Thus, the static feature represented by the point cloud and the dynamic feature represented by the MHI image are all included in the final fused descriptor.

4 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



w_1=0.1 w_1=0.2 w_1=0.3 w_1=0.4 w_1=0.5 w_1=0.6 w_1=0.7 w_1=0.8 w_1=0.9

Fig. 7 Descriptor fusion results with different weight values for the gesture "move straight"

Table 1 MSSIM index values obtained using different parameters for the gesture "move straight"

(w_1, w_2)	S1	S2	S3	S4	S5	S6	S7	S8
(0.1, 0.9)	0.2509	**0.4350**	0.2822	0.3761	0.2568	0.3402	0.3347	0.2938
(0.2, 0.8)	0.2168	**0.4061**	0.2389	0.3257	0.2260	0.2917	0.2714	0.2649
(0.3, 0.7)	0.2017	**0.3863**	0.2208	0.2950	0.2129	0.2668	0.2374	0.2448
(0.4, 0.6)	0.1915	**0.3708**	0.2108	0.2729	0.2012	0.2477	0.2159	0.2333
(0.5, 0.5)	0.1868	**0.3597**	0.2091	0.2616	0.1996	0.2380	0.2025	0.2279
(0.6, 0.4)	0.1846	**0.3485**	0.2068	0.2563	0.1983	0.2336	0.1923	0.2247
(0.7, 0.3)	0.1836	**0.3458**	0.2142	0.2559	0.1990	0.2351	0.1862	0.2251
(0.8, 0.2)	0.1779	**0.3527**	0.2242	0.2578	0.2009	0.2389	0.1802	0.2311
(0.9, 0.1)	0.1610	**0.3679**	0.2262	0.2639	0.1933	0.2407	0.1755	0.2265

straight, (3) left turn, (4) left turn waiting, (5) right turn, (6) change lane, (7) slow down, and (8) pull over. Figure 10 shows the eight gestures. Therefore, the goal of our method is to recognize the eight gestures of traffic police. For other gestures, once the fused descriptor of the gesture is obtained as a sample image included in the template database, the gesture can be recognized by following the same steps as mentioned below.

4.1 Mean structural similarity (MSSIM)

The mean structural similarity is based on the idea that a measure of change in structural information is a good approximation to perceived quality change. Generally, the extracted static feature image and the motion history image have strong structure characteristics. Meanwhile, the neighboring pixels have great correlation, which provides important information about the object structure in the viewing scene. Therefore, the MSSIM index can be used to measure the degree of similarity between the sample and the test image.

Specifically, the structural similarity (SSIM) index uses three separate comparisons of the local luminance (l), contrast (c), and structure (s) between the simple and test image. The SSIM is defined as [22]:

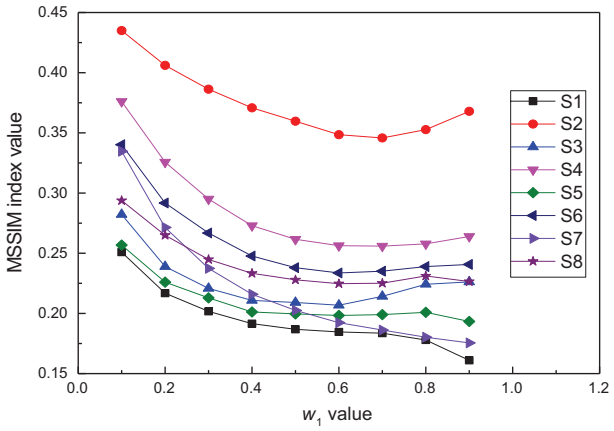


Fig. 8 MSSIM index values for the gesture "move straight"

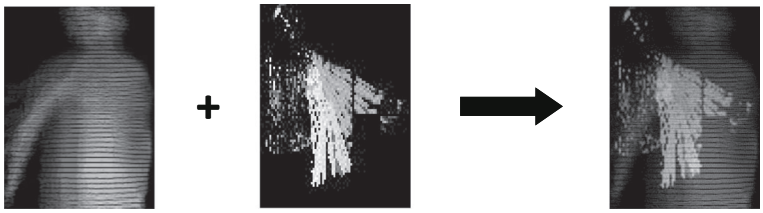


Fig. 9 Example of descriptor fusion result

$$\begin{aligned} \mathrm{SSIM}(x,y) &= [l(x,y)]^{\alpha}\cdot[c(x,y)]^{\beta}\cdot[s(x,y)]^{\gamma} \\ &= \frac{\left(2u_xu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(u_x^2 + u_y^2 + C_1\right)\left(\sigma_1^2 + \sigma_2^2 + C_2\right)} \end{aligned} \tag{7}$$

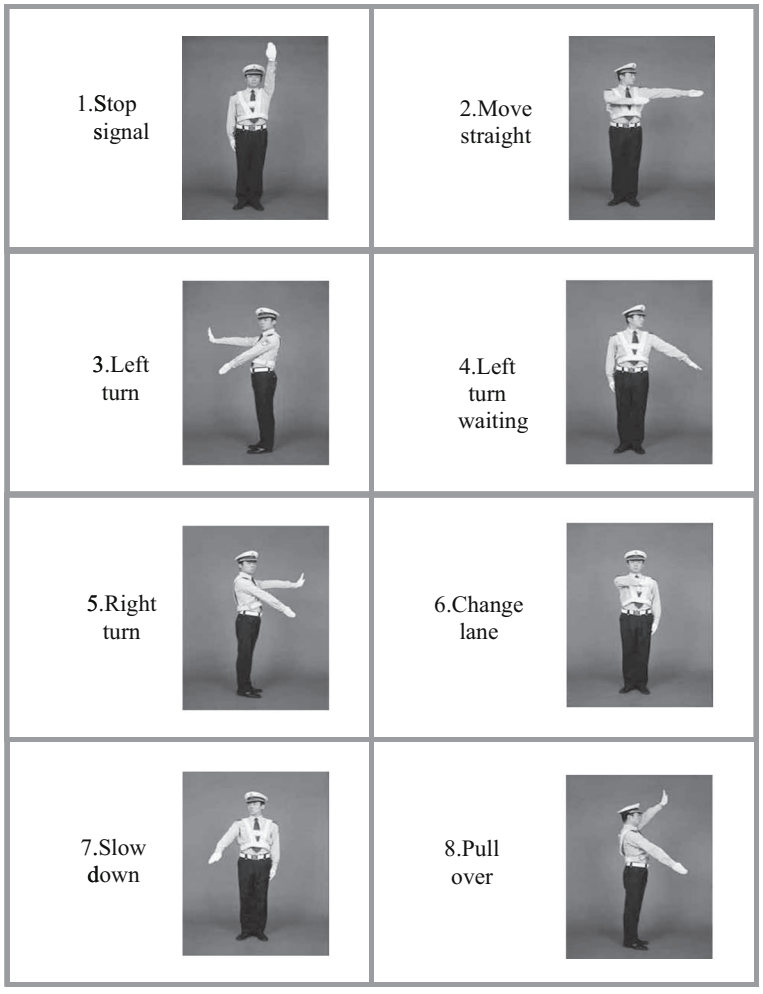


Fig. 10 Chinese traffic police gestures

where x and y represent two local image patches that are extracted from the simple and test images, respectively. α, β, and γ are the parameters used to adjust the importance of the three components. u_x and u_y are the mean value of the image patches x and y, respectively; σ_x and σ_y are the variance of x and y, respectively; σ_{xy} is the covariance between the image patches x and y. the constant C_1 is included to avoid instability when $u_x^2 + u_y^2$ is very close to zero:

$$C_1 = (K_1 L)^2 \tag{8}$$

where L is the dynamic range of the pixel values, and $K_1 < 1$ is a small constant. Similarly, $C_2 = (K_2 L)^2$, and $K_2 < 1$. In our experiment, we set $K_1 = K_2 = 0.5$ and $L = 100$ for all results reported in this paper.

In practice, one usually requires a single overall feature measure of the entire image. We use a mean structural similarity (MSSIM) index to evaluate the overall image feature.

$$\mathrm{MSSIM}(\mathrm{X}, \mathrm{Y}) = \frac{1}{M} \sum_{j=1}^{M} \mathrm{SSIM}\left(x_j, y_j\right) \tag{9}$$

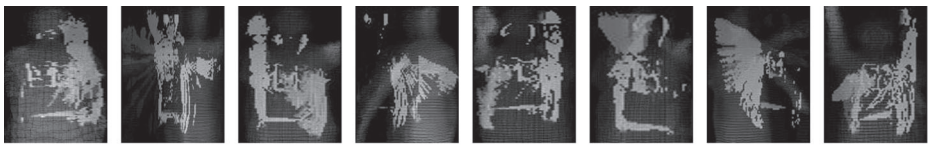
where X and Y are the sample and the test images, respectively; x_j and y_j are the image contents at the jth local window; and M is the number of local windows of the image. Since the higher the score value of MSSIM is, the more similar two images are, we can thus use the MSSIM index to recognize the police's gestures.

4.2 Gesture recognition using MSSIM

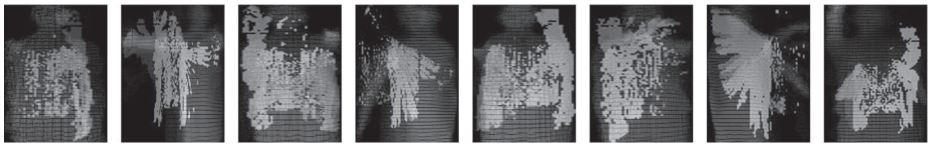
After obtaining the fused descriptors for sample and test image set, the value of MSSIM index is computed for classification. Supposed that traffic police gesture category c_i ($i = 1, 2, \ldots, p$) has N_i template training samples $B_j^{(i)} = (R_j^{(i)}, R_j^{(i)}, \ldots R_m^{(i)})$, ($j = 1, 2, \ldots, N_i$) $N = \sum_{i=1}^{p} N_i$ is the total number of the fused descriptor 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 fused descriptor of testing image B would be recognized, the MSSIM($B_j^{(i)}, B$) can be computed using Eq. (9). The MSSIM value is used to measure the similarity between B and template samples $B_j^{(i)}$. The following decision rule is taken to judge what categories of Chinese traffic police gesture the recognized gesture belong.

$$\text{If } S_m(B) = \max_{\substack{i \in (1,2,\ldots,p) \\ j \in (1,2,\ldots,N_i)}} \left\{ MSSIM_i\left(B_j^{(i)}, \mathrm{B}\right) \right\} and S_m(B)_{\text{index}} = c_m, \text{ then } B \in c_m. \tag{10}$$

A simple illustrative example is shown in Fig. 11. The descriptor fusion results of the eight standard police gestures constitute the template database, as shown in Fig. 11a. Eight descriptor fusion results obtained by real-captured test images are chosen for testing, as shown in Fig. 11b. From the MSSIM index obtained by the above method, we can deduce the classification results from the degree of similarity. Table 2 shows the MSSIM index value for Fig. 11, in which "S" stands for the sample image, and "T" stands for the test image. As can be seen in the table, the maximum values of MSSIM index are corresponding to the correct gesture categories. Besides, from Fig. 11, we can also see that some gestures, such as S1, S5 and S8 look similar to each other, but they are correctly recognized by using the MSSIM index shown in Tab. 2.



(a) Fused descriptors for sample image dataset



(b) Fused descriptors for test image dataset

Fig. 11 The fused descriptors for sample image dataset and test image dataset. **a** fused descriptors for sample image dataset, **b** fused descriptors for test image dataset

The high accuracy of the proposed method appears mainly due to two reasons: first, the fused descriptor is obtained by performing statistical analysis on the whole process of gestural motions, so the subtle differences of each gesture can be reflected in the final fused descriptor. Although some fused descriptors look very similar, there still exist some differences in the subtle features (e.g. pixel position, intensity, etc.) of the fused descriptor. This outcome constitutes one of the method's key advantages; second, the MSSIM indexing approach is a particular implementation of the philosophy of structural similarity from an image formation point of view [22], and the index can well measure the subtle differences in image structure since the index is proposed based on the degradation of structural information. Therefore, even if some gestures look much the same, these gestures can be correctly recognized by using the proposed method.

5 Experimental results

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, and also provide a way to deal with the situation when the arms of traffic police are perpendicular to the image plane since five gestures in the eight standard gestures of Chinese traffic police have such a situation. The police gesture recognition for moving vehicles is our research emphasis in the future. Besides, the proposed method just like most existing methods [10, 15] which recognizes police's gestures under indoor conditions since the gesture recognition in a real scenario is very difficult.

To test the proposed contribution, we measure gesture recognition of the Chinese traffic police while the fused descriptors were automatically computed by our proposed method. The assumption the proposed algorithm makes is that traffic police are seen approximately from a frontal viewpoint. In our experiment, the image dataset used for comparative study and qualitative evaluation was recorded in a controlled setting with no camera motion. The dataset contains the eight standard categories of gestures. There were 9 actors, including 1 police officer and 8 students, performing each gestures three times, resulting in about 72 video sequences in total. The 8 student actors were chosen from among voluntary undergraduate police management major students. All of them have knowledge on the gesture system of the

Table 2 MSSIM index value used for gesture recognition

	S1	S2	S3	S4	S5	S6	S7	S8
T1	**0.2871**	0.2213	0.2516	0.2211	0.2251	0.2407	0.2218	0.2037
T2	0.1868	**0.3597**	0.2091	0.2616	0.1996	0.2380	0.2025	0.2279
T3	0.2196	0.1890	**0.2440**	0.1760	0.2056	0.2172	0.1991	0.2023
T4	0.2409	0.2637	0.2377	**0.3091**	0.1966	0.2326	0.1996	0.2056
T5	0.2563	0.1988	0.2227	0.2118	**0.2993**	0.2039	0.1991	0.2040
T6	0.2497	0.2457	0.2625	0.1985	0.2216	**0.2752**	0.2462	0.2041
T7	0.1986	0.2009	0.2488	0.1556	0.2231	0.2710	**0.3548**	0.2140
T8	0.2360	0.2474	0.2761	0.2075	0.2700	0.2770	0.2490	**0.3362**

Chinese traffic. Prior to the study, each student actor was tested for proper gesture performance ability judging by the police officer, and those who failed the test did not participate in the evaluation.

5.1 Comparative study

Some representative human pose estimation algorithms were compared, such as max-covering scheme method [2] and stickman model method [6] with our proposed method. Since gestures can be recognized from the pose information of a human body, thus human pose estimation is the foundation of the gesture recognition. The reasons for choosing the above two pose estimation methods were that the former is proposed specifically for the gesture recognition of Chinese traffic police, and the latter is recognized as one of the most effective ways to estimate human pose. Besides, the CALVIN research group [6, 20] that proposed the stickman model method has many years' experiences in human pose estimation. The group has not only published many high-level papers in this field [4–6], but also developed a software named "2D articulated human pose estimation software v1.22" [20] based on their methods. The experimental results demonstrate that the method [6] is superior to other human pose estimation methods in term of dealing with the challenging uncontrolled imaging conditions, such as cluttered background, dark lighting, low contrast, and motion blur, etc. Therefore, we choose the above two methods for comparison with our proposed method. However, in our experiment, we find that there is no method can solve all the problems, and each method has its own drawbacks.

5.1.1 Comparison to max-covering scheme method

The max-covering scheme method [2] recognizes gestures by upper-body-part detection with a five-part body model. The five-part body model includes the torso, upper arms, and forearms. For this method, simple box detectors are used to find arm candidates. According to the max-covering scheme, only the candidate that completely cover the foreground pixels is kept, and the arm candidate is represented as a rectangle with a start side and an end side, as shown in the second row of Fig. 12.

Applying the five-part body model into the arm location, gesture recognition becomes simple and effective. However, since the five-part body model is a kind of two-dimensional model, it may not work for some particular gestures. When the police's arms are perpendicular

to the image plane, the five-part body model is invalid. The max-covering scheme method will underestimate the position of arms for these gestures, such as the gestures in Fig. 12. Our proposed method to recognize gestures was successful in locating the police's arms in the sequences. The gesture feature images perfectly matched the ground truth, as shown in the third row of Fig. 12. Thus, high recognition rate can be ensured by using the proposed method.

5.1.2 Comparison to stickman model method

For the stickman model method [6], 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. 13.

For our comparison with the stickman model method, upper-body-part detection was performed on the original frames shown in the first row of Fig. 12. As can be seen in the Fig. 12, the stickman model method was not able to locate the arms with the same efficiency as the proposed method. The reason for this is that the prior locations and the appearance transfer mechanism [9] used in the stickman method during body-part detection require a training stage, which is hard to satisfy for all kinds of person and environments.

We notice that the error will increase when the segment is far from the root joint in the six-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 stickman model method deviate from the ground truth, the estimated gesture category will deviate from the ground truth accordingly. Therefore, just like the max-covering scheme method, the correct recognition rate for the stickman method will be also very low no matter what recognition method is used.

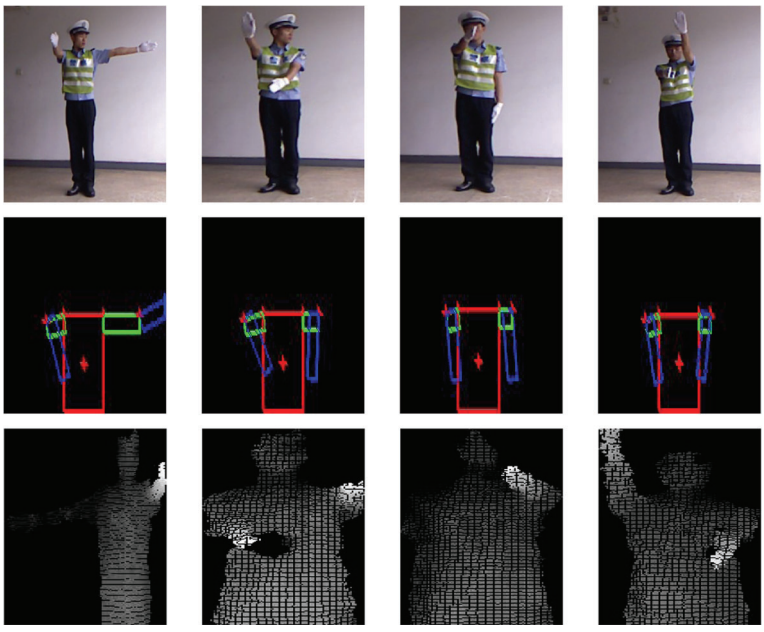


Fig. 12 Comparison between the arm location results of the max-covering scheme method and gesture feature images of the proposed method: The first row shows the orignal frames. The second row shows the arm location results of the max-covering scheme method. The geature feature images of the proposed method are shown in the last row

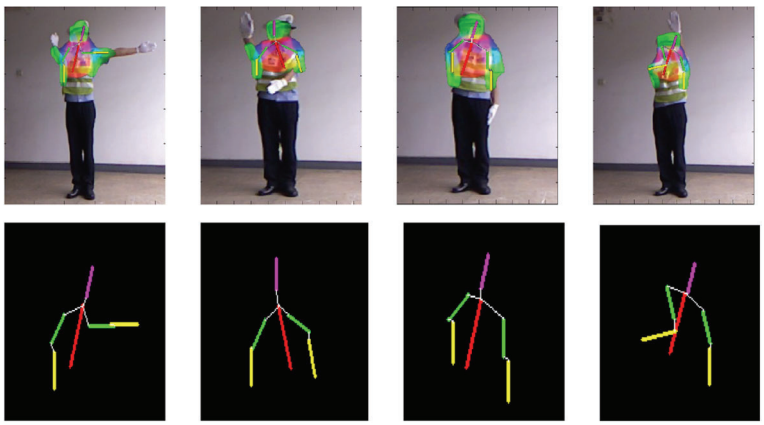


Fig. 13 The arm location results of the stickman model method for the original frames in Fig. 12. The first row shows the soft-labeling of pixels to body parts. *Red* indicates torso, *green* upper arms, *blue* lower arms and head. The second row shows the stickman representation of pose, obtained by fitting straight line segments to the segmentations in the first row. For enhanced visibility, the lower arms are in *yellow*, the upper arms are in *green* and the head is in *purple*

5.2 Qualitative evaluation

For qualitative evaluation, we mainly answer two questions. The first question was whether combing the static and dynamic descriptors really helps to improve the recognition rate of the police gestures. The second question was which pose estimation methods are most effective for recognizing police gestures: max-covering scheme method [2], stickman model method [6] or the proposed method.

5.2.1 Does combining static and dynamic descriptors really help?

The first question was whether combining static and dynamic descriptors helps to improve recognition performance. To determine this, we compared recognition performance under three conditions: static descriptor only (SDO), dynamic descriptor only (DDO) and static and dynamic descriptors (SDD), i.e., SDO contained only static features, DDO contained only dynamic features, and SDD contained both static and dynamic features. Figure 14 shows eight gesture comparisons of the three conditions (SDO, DDO and SDD).

As can be seen in Fig. 14, the performance difference between SDO and SDD was significant for 2 gestures (#3 and #5) where the arm pose plays an important role in defining these gestures (see Fig. 10). Since the difference between the feature image of "left turn (#3)" and "change lane (#6)" is very small, it is very hard to differentiate the gestures just by static descriptor. The gesture pair "right turn (#5)" and "left turn (#3)" share the same problem. For DDO and SDD, the difference is especially obvious for gesture pair #2 and #8. That's because the discrepancy between "move straight (#2)" and "change line (#6)" is tiny from the view of the MHI image, the same is for the gesture pair "pull over (#8)" and "right turn (#5)". Our result indicated that using static and dynamic descriptor together on these 4 gestures (#2, #3, #5, and #8) can achieve higher accuracy; for the other 4 gestures where there were slight difference, but none was significant.

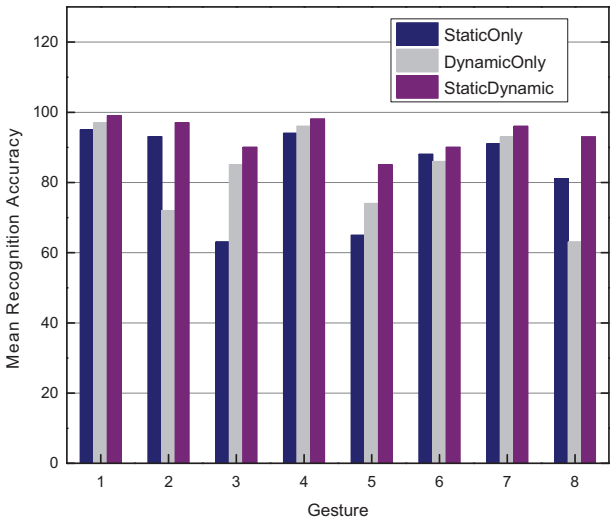


Fig. 14 Eight gesture comparisons of StaticOnly, DynamicObly and StaticDynamic. (gesture numbers: *#1*, stop; *#2*, move straight; *#3*, left turn; *#4*, left turn waiting; *#5*, right turn; *#6*, change lane; *#7*, slow down; *#8*, pull over)

5.2.2 Which human pose estimation methods are most effective?

Various human pose estimation methods have been explored in gesture recognition research, but there is no clear sense as to which method is most effective. In response, we compared the recognition accuracy using various human pose estimation methods, including max-covering scheme method, stickman method and the proposed method.

We perform the three methods by executing MATLAB on a PC with a 3.00 GHz Intel Pentium Dual-Core Processor. The gestures in the video include the eight police gestures, and each gesture appears in a test video clip. For the max-covering scheme method and stickman method, the Gabor-based 2DPCA [2, 8] is adopted to recognize police gestures in the testing video. For the proposed method, the mean structural similarity (MSSIM) index is adopted to recognize gestures with the fused descriptor. All the video clips are captured approximately from a frontal viewpoint. Each time a police is spotted, the closest gesture class will be assumed. Figure 15 presented the correct rate of all the three methods. As can be seen, all eight gestures can be recognized using these methods. However, the max-covering scheme method and stickman method share the common limitation of most pose estimation methods—the two-dimensional model used in these methods may be invalid for some particular gestures.

From Fig. 15, we can order the three methods in decreasing order with respect to average correct rate: proposed method, max-covering scheme method and stickman method. Stickman method relies on a training stage during body-part detection. This requirement limits its application for all kinds of person and environments. For max-covering method, the 5-part body model can't accurately locate the position of police's arms for some particular gestures are its biggest limitation. Thanks to the good fused descriptor which combines both static feature and dynamic feature, the proposed method performs significantly better than the other two methods.

Besides, experimental results also show that, some gestures (e.g., "stop," "move straight," "left turn waiting", "slow down") can achieve a high recognition rate, whereas other gestures ("left turn," "right turn", "change lane", "pull over") have a relatively lower rate for the max-covering scheme

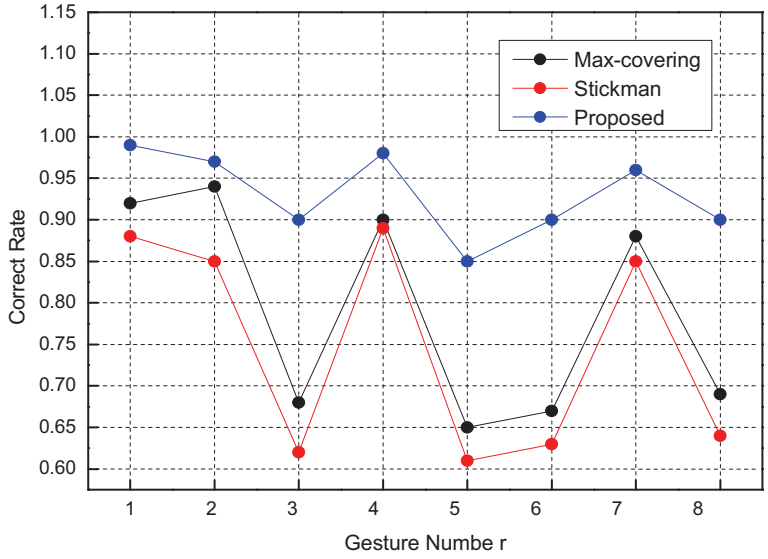


Fig. 15 Correct rate for different human pose estimation methods. (gesture numbers: #*1*, stop; #*2*, move straight; #*3*, left turn; #*4*, left turn waiting; #*5*, right turn; #*6*, change lane; #*7*, slow down; #*8*, pull over)

method and stickman method. The reason of the low recognition rates for the two methods may be mainly due to the difficulty in accurately locating the arms in previous steps. On the contrary, the proposed method can achieve a good recognition rate for all the eight gestures, as shown in Fig. 15. Thus, we can deduce that for all the features that used in the three methods, a combination of static descriptor and dynamic descriptor was the most informative feature for this task.

5.2.3 Computation times

In the experiment, the frame resolution of the input testing video is 640 × 480. For the max-covering schemem method, the computational time for detecting and recognizing the police gestures is approximately 14 s per frame in the MATLAB environment. The stickman model method is a semi-automatic method, requiring manually drawing a bounding box around the hand and shoulders of a policeman in each frame image as the algorithm input. After locating the position of the bounding box, the algorithm takes less than 2 s to output a set of line segments indicating the body part information for each frame image, and needs another 5 s to recognize the police gestures using the Gabor-based 2DPCA in the MATLAB environment. The main difference between the above two methods is that the stickman model method is a supervised learning approach that predicts the body part information with a training set of annotated images, whereas the max-covering scheme method is an unsupervised learning approach that uses 5-part body model to detect police's arm location without using a training set. The proposed method is also an unsupervised learning approach. It takes about 9 s to extract the gesture feature, and less than 1 s to recognize gestures using the MSSIM index. Although the time spends on obtaining the fused descriptor seems a litter longer, the proposed method can handle well with the situation when the arms of traffic police are perpendicular to the image plane, while the other two methods may be invalid in this situation. This contribution is very important since five gestures in the eight standard gestures of Chinese traffic police have such a situation.

6 Discussion

In this section, we mainly discuss some critical issues, such as technique challenges, limitations, and possible solutions, that relate to the technique contribution of the proposed method.

Gesture recognition of Chinese traffic police faces many technique challenges: a) the distance of a Kinect camera to the policeman would be too far to get reasonable accuracy; b) the unpredictable environment, such as changing weather and changing lighting conditions, as well as a lot of occlusions and reflections due to the car lights; c) cluttered background, such as high-density crowds and vehicles in the scene; d) clear gestures information is hard to get with the moving vehicles.

Although some measures have been taken to solve a part of the above problems using a single RGB camera, the recognition rate for some police gestures is very low. For example, to detect traffic police in a complex traffic scene, in our previous work the police's torso and arms are first extracted from the complex scene as the foreground region by using dark channel prior and kernel density estimation [2]. The method can be used in an uncontrolled real scenario setting, but as a result of the lack of any depth information, the human body model constructed in the method is a 2D model. Thus, it is hard to deal with the situation when the arms of traffic police are perpendicular to the image plane. Therefore, a new gesture recognition method is proposed in this paper to cope with this problem. However, the proposed method also has some limitations: a) a relatively stationary background of video sequence is required for the proposed method; b) the police are required to be in focus, visible, and not blurry. If the background is very complex or the vehicle is moving fast, the image may very well be noisy or blurry in that case, and our method will underestimate the foreground for the police; c) even though high quality of input raw data is guaranteed, for example, the proposed features (accumulated temporal features) might not be robust in background clutter with dynamic motions. 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.

In the future, we intend to investigate the following possible solutions to enhance the flexibility of the proposed algorithm: a) incorporating some operations, such as fog and raindrop removal, low-lighting enhancement, background removal, and camera shake removal, etc., into the preprocessing stage; b) studying the robust and effective human body detection methods for both RGB data and depth data in an unpredictable traffic scene; c) we also believe that the gesture recognition task would benefit from the advanced color-depth sensors which may be more proper for outdoor applications.

7 Conclusion

In this paper, we have proposed a very simple but effective algorithm for recognizing the gestures of Chinese traffic police. Key features of our proposed method are the fusion of the extracted static and dynamic descriptors and the use of a mean structural similarity index 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 adopts a 2.5D gesture model and a motion history image to extract good static and dynamic features. Thus, the gestures can be recognized even if the police's arms are perpendicular to the image plane.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (No. 61502537, 91220301), China Postdoctoral Science Foundation (No. 2014 M552154), Hunan Planned Projects for Key Scientific Research Funds (No. 2015WK3006), Postdoctoral Science Foundation of Central South University (No. 126648).

References

1. Bradski G, Davis J (2000) Motion segmentation and pose recognition with motion history gradients. In: Proceedings of IEEE Workshop on Applications of Computer Vision, pp 174–184
2. Cai ZX, Guo F (2015) Max-covering scheme for gesture recognition of Chinese traffic police. Pattern Anal Applic 18(2):403–418
3. Eichner M, Ferrari V (2009) Better appearance models for pictorial structures. In: Proceeding of British Machine Vision Conference, London, UK, pp 1–11
4. Eichner M, Ferrari V (2012) Human pose co-estimation and applications. IEEE Trans Pattern Anal Mach Intell (PAMI) 34(11):2282–2288
5. Eichner M, Marin-Jimenez M, Zisserman A, Ferrari V (2012) 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. Int J Comput Vis (IJCV) 99(2):190–214
6. Ferrari V, Marin-Jimenez M, Zisserman A (2008) Progressive search space reduction for human pose estimation. In: Proceeding of IEEE Conference on Computer Vision & Pattern Recognition, Anchorage, AK, pp 1–8
7. Guo F, Cai ZX, Tang J (2011) Chinese traffic police gesture recognition in complex scene. In: Proceeding of the 2011 International Joint Conference of IEEE FCST-11, Los Alamitos, USA, pp 1505–1511
8. Guo F, Tang J, Cai ZX (2013) Automatic recognition of Chinese traffic police gesture based on max-covering scheme. Adv Inf Sci Serv Sci 5(1):428–436
9. 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
10. Johnson S, Everigham M (2011) Learning effective human pose estimation from inaccurate annotation. In: Proceeding of IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, USA, pp 1465–1472
11. Kang H, Lee CW, Jung K (2004) Recognition-based gesture spotting in video games. Pattern Recogn Lett 25(15):1701–1714
12. Le QK, Pham CH, Le TH (2012) Road traffic control gesture recognition using depth images. IEEK Trans Smart Process Comput 1(1):1–7
13. Liu JG, Luo JB, Shan M (2009) Recognizing realistic actions from videos 'in the wild'. In: Proceeding of IEEE Conference on Computer Vision & Pattern Recognition, Miami, FL, pp 1996–2003
14. Sapp B, Jordan C, Taskar B (2010) Adaptive pose prior for pictorial structures. In: Proceeding of IEEE Conference on Computer Vision & Pattern Recognition, San Francisco, USA pp 422–429
15. Singh M, Mandal M, Basu A (2005) Visual gesture recognition for ground air traffic control using the Radon transform. In: Proceeding of IEEE/RSJ International Conference on Intelligent Robots & Systems, Edmonton, Canada, pp 2586–2591
16. Smisek J, Jancosek M, Pajdla T (2011) 3D with Kinect. In: Proceedings of the 2011 I.E. International Conference on Computer Vision Workshops, Barcelona, Spain, pp 1154–1160
17. Song Y, Demirdjian D, Davis R (2011) Tracking body and hands for gesture recognition: NATOPS aircraft handling signal database. In: Proceeding of IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, Santa Barbara, CA, pp 500–506
18. Suau X, Casas JR, Ruiz-Hidalgo J (2011) Real-time head and hand tracking based on 2.5D data. In: Proceedings of the 2011 I.E. International Conference on Multimedia and Expo, Barcelona, Spain, pp 1–6
19. Tang J, Luo J, Tjahjadi T, Gao Y (2014) 2.5D multi-view gait recognition based on point cloud registration. Sensors 14:6124–6143
20. Visual Geometry Group (2015) 2D articulated human pose estimation software v1.22, http://groups.inf.ed.ac.uk/calvin/articulated_human_pose_estimation_code/. Accessed 15 May 2015
21. Yuan T, Wang B (2010) Accelerometer-based Chinese traffic police gesture recognition system. Chin J Electron 19(2):270–274

22. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
23. Zhou Z, Li ST, Sun B (2014) Extreme learning machine based hand posture recognition in color-depth image. In: Proceedings of Chinese Conference on Pattern Recognition, pp 1–10
24. Zhu Y, Fujimura K (2010) A Bayesian framework for human body pose tracking from depth image sequences. Sensors 10:5280–5293
25. 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

Fan Guo received the B.S. degree in computer science and technology from the Central South University (CSU), Changsha, China, in 2005, and the M.S. degree and the Ph.D. degree in computer application technology from CSU, Changsha, China, in 2008 and 2012, respectively. Currently, she is a postdoctoral fellow and Lecturer with the School of Information Science and Engineering, CSU. Her main research interests include image/video processing, pattern recognition and virtual reality.

Jin Tang received the B.S. degree and the M.S. degree from Peking University, Beijing, China, in 1987 and 1990, respectively, and the Ph.D. degree in pattern recognition and intelligence system from Central South University (CSU), Hunan, China, in 2002. He is currently a Professor in the School of Information Science and Engineering, CSU, Changsha. His current research interests are focused on image processing, pattern recognition and computer vision.