ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 644-649
doi: 10.18178/wcse.2016.06.113

An Improved Recognition Method of Traffic Gestural Command Based on Skeleton Data

Fan Huang, Qian Huang and Chongwen Wang [+]

School of Software Engineering, Beijing Institute of Technology, Beijing, 100081, China

Abstract. This paper proposes an improved method based on skeleton data to increase the recognition efficiency and recognition rate of traffic gestural command. First, Kinect is used to obtain the coordinate data of skeleton nodes. Then the cosine value of deflection angle is extracted as feature for training and testing. Finally, DTW (Dynamic Time Warping) algorithm is improved by deflection weighting and sample filtering to recognize the given samples. Experiment shows that the average recognition rate of the method in this paper is up to 97%.

Keywords: traffic gestural, skeleton data, gesture recognition, DTW.

1. Introduction

Using gesture command to direct the traffic, in some special environment, such as: sudden accident, snow fog, night driving, can effectively solve the traffic congestion problem. To help drivers have a better understanding of these commands, a group of researchers is focusing on automatic recognition of traffic gestural command [1]. According to different gesture acquisition modes, the recognition methods can be divided into two kinds.

The first kind is to obtain gesture data from motion capture devices like data gloves. Kim and Chien [2] used data glove technique to obtain the three-dimensional positions of the hand, from which the hand trajectory was derived. Xiao Qian and Yang Ping [3] bound a number of sensors which have high accuracy of measurement on the police’s hands to discern the command gesture of the police. These methods received good results of gesture recognition. However, measure devices with are often very expensive. In addition, wearing a certain number of sensors may make the users feel uncomfortable.

The second kind is getting data through cameras or other video capture devices. Image segmentation, feature extraction, feature recognition will be used to complete a gesture recognition. Amor, Su and Srivastava [4] studied the problem of classifying actions of human subjects using depth movies generated by Kinect. They obtained the skeletal shapes of actions from Kinect and used the sequences of shapes to classify human actions. Li [5] proposed a recognition algorithm based on contour deletion, which successfully recognized the traffic police gestures. Liu and Shang [6] extracted the Euclidean distance, which also has good results in recognition. The use of machine vision based approaches can greatly reduce limitation on action of gestures. However, methods based on machine vision often need to process large amounts of data, especially the training process, which reduces the efficiency of recognition.

This paper presents an improved method to raise the accuracy and efficiency of gesture recognition. It improves the feature extraction method and the matching algorithm to simplify the recognition process.

The contributions of this paper are summarized as follows:

[+] Corresponding author. Tel.: +86 18600021476.
E-mail address: wcwzzw@bit.edu.cn.

- We choose cosine value of deflection angle as the extracted feature. This feature can eliminate differences among human bodies and is proved to be suitable in the experiments.
- We increase the accuracy of recognition by giving each skeletal node a weight. Each node contributes differently to the trajectory of gesture. The weight of each node can make the model exacter.
- We increase the efficiency of recognition by giving each gesture an offset. The offset in this paper can filter out samples which have less overlap with test gesture..

2. Pre-processing

We use the 3D coordinates (x, y, z) which provided by Kinect to represent the skeletal joints. As shown in Fig. 1, z-axis has the same orientation with Kinect. Y-axis is vertically upward and x-axis horizontally towards left (watching from Kinect). The unit of coordinate system is meter.

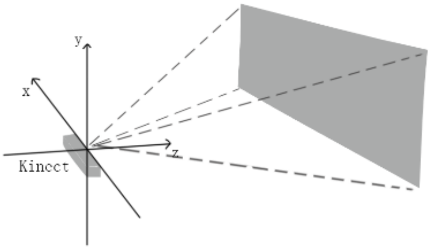

Fig. 1: 3D coordinate system according to Kinect.

Min-max, also known as the deviation of standardization, is mainly used to make linear transformation on the original data, thus the outcome values are mapped to interval [0-1]. The equation (1) can make the conversion.

$$x^* = (x - \min)/(max - min) \tag{1}$$

The 8 kinds of traffic gestural commands are unrelated individuals, separated and independent from each other. Based on this characteristic, we can consider each of the gestures as a recognition object. As for a specific gesture, the maximum and minimum value of all the sample data can be determined, thus min-max standard method is used for data normalization, as shown in equation (2).

$$\begin{cases} x^* = (x - x_{min})/(x_{max} - x_{min}) \\ y^* = (y - y_{min})/(y_{max} - y_{min}) \\ z^* = (z - z_{min})/(z_{max} - z_{min}) \end{cases} \tag{2}$$

3. Gestural Feature Extraction

The basic characteristics of gesture track are velocity, position, distance and angle. J. Singha and R. H. Laskar [7] used six features in their paper: location, orientation, velocity, position, self co-articulated, ratio and distance. The combination of these features has good results in recognition. However, in this paper, we aim to propose a simplified method to recognize the traffic command gestures so that we want to choose only one feature to extract. Liu Yang and Shang Zhaowei [6] raise a feature extraction method in their paper, using the Euclidean distance between each node and the shoulder center node as the feature. The intention is to transform the 3D feature into one-dimensional feature, it’s a good idea but it can’t eliminate effects caused by differences among human bodies such as height and length of arms.

The analysis of 8 traffic gestures shows that these are very standard gestures, and are strictly restricted by the behavior essentials, in which requirements on angle appear frequently. We can make a conclusion that angle can be a measurement for the command gestures. So in this paper, we use cosine value of the deflection angle as the feature of the traffic police gestures.

In three-dimensional space, two vectors determine an angle. We choose a vector points from the joint node to the shoulder center node, the other one is a unit vector pointing vertically down. The equation to calculate cosine angle made by these two vector is shown in equation (3).

$$\cos\theta_i^t = (op_i^{t'} * \overrightarrow{os})/(\left|op_i^{t'}\right| * |\overrightarrow{os}|) \quad (3)$$

The traffic police gesture command is a dynamic process, in which the node will produce an angle cosine in every frame captured, so the trajectory formed by all the cosine value produced by a node can be defined as the node's motion track. Definition of the motion track is shown in equation (4).

$$\theta_i = \cos\theta_i^1, \cos\theta_i^2, \cos\theta_i^3, \ldots, \cos\theta_i^t \quad (4)$$

The 8 traffic command gestures are determined by the motion tracks of 4 nodes shown in Fig. 2: left hand (LH), left elbow (LE), right hand (RH), and right elbow (RE).

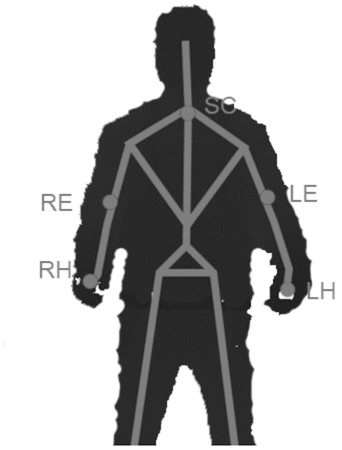

Fig. 2: Kinect skeleton data.

4. Gesture Recognition

Vision based hand gesture recognition techniques can be further classified under static and dynamic gestures. The mainly used algorithms to recognize dynamic gestures are Hidden Markov Models (HMM) and Dynamic Time Warping (DTW) [8]. Camona and Climent [9] evaluated the recognition results of these two algorithms and proposed the advantages of DTW algorithm. So we choose DTW algorithm.

4.1. Principle of DTW

DTW (Dynamic Time Warping) is used to capture the similarity of two time-related sequences [10]. This paper takes motion tracks of bone nodes as feature, so the smallest unit of recognition is the track of each node. The following assumptions are made based on this premise.

Assume that there is a standard gesture command W, name its left hand joint node as LH. Its motion track, also called reference template is a vector of m-dimension, like R=$(r_1, r_2, r_3, \ldots, r_m)$, each of these components is the cosine angle between the vector from LH node to the SC node and the unit vector pointing vertically down, besides m is frames of the standard action. Then there is a test command gesture W, the motion track of its LH node which is also called the test template is T, T=$(t_1, t_2, t_3, \ldots, t_n)$. Each component of T is also the cosine angle between the two vectors and n is frames of this test command gesture.

From the above we can see that each component of R and T belongs to the same category, so the similarity or distance can be calculated by DTW algorithm. The task of DTW is to find an optimal matching between R and T which means the sum of distances between corresponding elements comes to a minimum value. Use $\emptyset = (\emptyset_R, \emptyset_T)$ to represent the optimal matching between R and T, as shown in equation (5) and (6). So the DTW between R and T is defined as formula (7).

$$\emptyset_R = \left(\emptyset_R^1, \emptyset_R^2, \emptyset_R^3, \ldots, \emptyset_R^k\right), \emptyset_R^k \in \mathrm{R}, 1 \ll \mathrm{k} \ll \mathrm{m} \quad (5)$$

$$\emptyset_T = \left(\emptyset_T^1, \emptyset_T^2, \emptyset_T^3, \ldots, \emptyset_T^k\right), \emptyset_T^k \in \mathrm{T}, 1 \ll \mathrm{k} \ll \mathrm{n} \quad (6)$$

$$\emptyset_T = \left(\emptyset_T^1, \emptyset_T^2, \emptyset_T^3, \ldots, \emptyset_T^k\right), \emptyset_T^k \in \mathrm{T}, 1 \ll \mathrm{k} \ll \mathrm{n} \quad (7)$$

Among them, d(i, j) is the distance between r_i in R and t_j in T. Therefore, the solution of finding the optimal matching can be transformed. First, build a m*n sized two-dimensional matrix M, m is the length of

R sequence, n is the length of T sequence, elements of M defined as δ_{ij} are the DTW between $\mathrm{R} = (r_1, r_2, r_3, \ldots, r_m)$ and $\mathrm{T} = (t_1, t_2, t_3, \ldots, t_n)$. To calculate the minimum value δ_{min}, all the possible matching can be exhausted, but the time complexity will be exponential. As the sequence of gesture is sequential, δ_{ij} or M[i, j] can be defined by the sum of d(i, j) and the minimum among M[i-1, j-1], M[i-1, j] and M[i, j-1]. Formula (8) can be used in order to reduce the complexity of the algorithm.

$$\delta_{i,j} = \mathrm{d}(i,j) + \min\{\delta_{i-1,j-1}, \delta_{i-1,j}, \delta_{i,j-1}\} \tag{8}$$

4.2. DW-DTW

To recognize which command the captured data is, each of the nodes' DTW value should be calculated and then added together. The final DTW value determines the matching gesture.

From the analysis of traffic command gestures we can clearly know that the motion states of each node in different gestures are different. For example, in the stop sign, only node LH and LE actually move and the other nodes are relatively motionless. Thus, the calculated DTW values of each node in different gestures contribute in varying degrees to the final classification for gestures. If we simply add up these DTW values, we can't obtain a high rate of identification although some of the gestures can be distinguished. The literature proposes a method to improve the recognition rate by giving different weights to different nodes. In this paper, we raise a deflection angle weighted dynamic programming algorithm (DW-DTW). For gesture G, the cosine sum of its node i's deflection angle is defined as formula (9).

$$D_i^g = \sum_{t=1}^{T} \left|\cos\theta_i^t - 1\right| \tag{9}$$

Among them, T is total frames of G. According to formula (9), the weight of node i can be defined as formula (10).

$$\mathrm{w}_i^g = \frac{\mathrm{D}_i^g}{\sum_{k=1}^{N} \mathrm{D}_k^g} \tag{10}$$

It's obvious that if a node of gesture G is always in a static state, the weight will be zero. Therefore, the final DTW value can be calculated by formula (11).

$$\delta = \sum_i^N w_i^g \delta_i \ (N = n) \tag{11}$$

Among them, w_i^g is weight of node i; δ_i is DTW value. By calculating the weight of each node in a gesture, contribution of each node is objectively shown, which improves the precision of gesture description.

4.3. DWF-DTWh

Calculation of DTW between T and S is comparatively time-consuming and is meaningless in some cases, for example there is almost no overlap between stopping gesture and turning left gesture. If calculate without limits, it will be a waste of computational resources thus reduce the efficiency of identification.

So this paper puts forward a method of deflection filter (DWF-DTW). This method proposes an angle based deflection value aw, defined as equation (12).

$$\mathrm{aw}(g) = \sum_i^p 10^i \times D_i^g \ (p = 7) \tag{12}$$

Among them D_i^g is the cosine value of node i's deflection angle in gesture g, defined in formula (9). 10^i is weight of node i, for example node 1 weights 10 and node 2 weights 100. $\mathrm{aw}(g)$ is the sum of 7 nodes. In formula (12), each node is given a different weight from others to avoid the offset among nodes, such as 2+8=6+4. After method of deflection filter, the range of contrast reduces to 2 or 3 gestures, so the improvement in efficiency is about 62.5%~75%.

5. Experiment and Result

5.1. Training

We asked 10 traffic police to act the 8 traffic commands. We extracted cosine value of deflection angle for node LH, LE, RH, RE. Then we calculate the aw(g) for each gesture according to formula (12). The results are shown in Table 1.

Table 1. The aw(g) Value for Each Gesture

Number	Gesture	aw
1	Stop command	11876.43
2	Go straight command	146905.33
3	Turn left command	218986.48
4	Change lane command	132075.91
5	Turn left waiting command	21454.12
6	Turn right command	54587.72
7	Slow down command	96594.58
8	Pull over command	155361.71

5.2. Experiment

To evaluate the performance of our proposed method, we asked 30 subjects to act the traffic commands. Then we captured images of size 640*480 with Kinect. Details of the experiment are as follows.

Experimental equipment: Kinect with a resolution of 640*480 for depth image.

Experimental object: 30 subjects with height 160-185 cm.

Experimental process: each subject acts randomly the command gestures 30 times.

The 30 subjects are grouped by their heights. The number of gestures acted by each group is shown in Table 2. In Table 2, number is the number of subjects in the group. G1, G2, … G8 are the 8 traffic gesture in Table 1.

Table 2. The Number of Commands Acted by Each Group

Group	Height	Number	G1	G2	G3	G4	G5	G6	G7	G8
1	160-165	7	27	28	26	23	28	26	24	28
2	165-170	6	21	20	25	26	23	19	24	22
3	170-175	7	24	29	27	22	23	29	27	29
4	175-180	7	26	26	24	25	27	24	28	30
5	180-185	3	12	10	12	13	9	11	13	10
Total number		30	110	113	114	109	110	109	116	119

5.3. Result

We implemented the HMM algorithm as the recognition method. The recognition results of our method and the HMM method are shown in Table 3.

Table 3. Recognition Results of HMM, DTW, DW-DTW, DWF-DTW

Recognition method	Test number	Correct number	Recognition Rate	Recognition time
HMM	900	572	63.6%	223 ms
DTW	900	698	77.5%	125 ms
DW-DTW	900	855	95%	136 ms
DWF-DTW	900	877	97.4%	43 ms

As Camona and Climent [9] have shown in their paper, the DTW algorithm has better results than HMM. The results also show that the improved DWF-DTW algorithm proposed in this paper has a better performance in the accuracy and efficiency than the original DTW algorithm. As for accuracy, the improved DTW methods DW-DTW and DWF-DTW have accuracy of about 95%, while the accuracy of original DTW algorithm is 77.5%. As for efficiency, the DTW and DW-DTW algorithm are time-consuming. After adding a filtering method, DWF-DTW algorithm is two times faster.

We also implemented the algorithm proposed by Liu Yang and Shang Zhaowei [6]. Table 4 shows the recognition results of Liu's method and Table 5 shows the results of our method.

Table 4. Recognition Results Based on Euclidean Distance

Group	Height	Number	Test number	Correct number	Recognition Rate
1	160-165	7	210	183	87.1%
2	165-170	6	180	156	86.7%
3	170-175	7	210	185	88.1%
4	175-180	7	210	201	95.7%
5	180-185	3	90	85	94.4%

Table 5. Recognition Results Based on Cosine Value of Deflection Angle

Group	Height	Number	Test number	Correct number	Recognition Rate
1	160-165	7	210	205	97.6%
2	165-170	6	180	176	97.8%
3	170-175	7	210	204	97.1%
4	175-180	7	210	204	97.1%
5	180-185	3	90	88	97.8%

From Table 4 we can see that, the method based on Euclidean distance has higher accuracy when the subjects' height are higher than 175cm. This is probably because our sample data is collected from 10 traffic police. They are about 180cm. Table 5 shows that our method based on deflection angle can eliminate the differences in human bodies.

It's clear that the improved method DWF-DTW raised in this paper has a better performance in recognition rate and recognition time. The recognition rate of this method is 97.4% and the recognition time is about 43ms.

6. Conclusion

We have proposed an improved recognition method of traffic gestural commands. According to the characteristics of captured Kinect skeletal data, we have presented a new feature-extraction method based on deflection angle which is determined by the vector from skeletal node to the center shoulder node and the unit vector pointing vertically down. This method greatly reduces effects caused by different body height and arms' length among people. In addition, we have improved the DTW algorithm by adding a sampling filter which can shorten the recognition time. Experimental result shows that using the method in this paper can increase not only the efficiency but also the accuracy of gesture recognition. In our future work, we will try to use more classifiers such as ANN, SVM and kNN to make a further comparison.

7. References

[1] W. J. Song, M. Y. Fu, and Y. Yang, Y. Chen. Recognition Method of Traffic Police and Their Command Action Based on Kinect. *Proceedings of the 33rd Chinese Control Conference*, Nanjing, 2014.

[2] Kim, and Chien. Analysis of 3D hand trajectory gestures using stroke based composite hidden Markov models. *Appl. Intell.*. 2001, **15**(2): 131-143.

[3] Q. Xiao, and P. Yang. A Gesture Recognition Method Based on MEMS IMU. *Chinese Journal of Sensors and Actuators*. 2013, **26**(5): 611-615.

[4] B.B.Amor, J.Y.Su, and A.Srivastava. Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2015.

[5] W. J. Li. Gesture Recognition for Traffic Control based on Thinning Algorithm and Template Matching. *M.S. thesis*, 2011.

[6] Y. Liu, and Z. W. Shang. Traffic Gesture Recognition Based on Kinect Skeleton data. *Computer Engineering and Applications*. 2015, **51**(3): 157-161.

[7] J. Singha, and R. H. Lasker. Self Co-articulation Detection and Trajectory Guided Recognition for Dynamic Hand Gestures. *Iet Computer Vision*, 2015.

[8] S. S. Rautaray, and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*. 2015, **43**(1): 1-54.

[9] J. M. Camona, and J. Climent. A Performance Evaluation of HMM and DTW for Gesture Recognition. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. 2012, **7441**: 236-243.

[10] D. J. Berndt, and J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. *Working Notes of the Knowledge Discovery in Databases Workshop*. 1994: **10**(16): 359-370.

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, et al. Real-Time Human Pose Recognition in Parts from Single Depth Images. *Communications of the Acm*. 2011, **411**(1): 1297-1304.

[12] P. Ronald. Vision-based Human Motion Analysis. *Comput Vision and Image Understanding*. 2007, **108**: 4-18.

[13] T. Yuan, B. Wang. Accelerometer-based Chinese Traffic Police Gesture Recognition System. *Chinese Journal of Electronics*. 2010, **19**(2): 270-274.