Vision-Based Gesture Recognition Referring to Human Structure

Shihai Xu [a], Binkai Zhang [b], Xiao Liang [c] and Jinjin Zheng [d]

PMPI, University of Science and Technology of China, Hefei 230027, China

[a]haizi@mail.ustc.edu.cn, [b]bkzhang@mail.ustc.edu.cn, [c]xiao@mail.ustc.edu.cn, [d]jjzheng@ustc.edu.cn

Keywords: Gesture Recognition, Body Structure, Vision, KNN, Camshaft.

Abstract. This article proposed a new method which introduces human body structure into gesture recognition besides the traditional vision information. By establishing the target's skeleton model, the system could rule out background interference quickly from the real-time frames, then locate the target's joints accurately, and track the gestures automatically. This method applies a common monocular webcam for image input instead of expensive depth cameras or complex binocular cameras. It can build a vivid human model and simulate the target's gestures well. After a modified KNN classification which has self-learning ability, the relationships between these joints are analyzed, and the corresponding postures and movements are identified. This method avoids global scanning in each frame and simplifies the classification, thus the computation decreases sharply. The results show that the method has a good tolerance of complex background, it can solve the target missing well.

Introduction

Human-Computer Interaction (HCI) has become one of the major technologies today. The core issue of HCI is Vision-Based Interface (VBI), wherein gesture recognition is an important part [1].

Gesture, as a natural and intuitive expression, contains a lot of information in line with human habits. Using hands as input devices and controlling other facilities have drew more and more attention for the advantages of convenience, non-contact, low cost and large information.

There are four main problems in most existing methods based on vision [2-5]: 1. Static background; 2. Excluding non-skin area incompletely; 3. Positioning wrist vaguely; 4. Unintelligent recognition process [6]. In some systems, additional information, like depth [7-8], mark, which requires complicated equipment, is used to judge the target's intension. In this article, human body structure are introduced and compared with the vision information so as to improve the recognition accuracy, speed and adaptability to different scenes.

Gesture recognition system

In the proposed method, gesture recognition system is divided into four main parts: 1.Image acquisition, this module uses a simple webcam for image input; 2. Image preprocessing; 3. Skeleton modeling, the system has set two modes dealing with different scenarios: body skeleton modeling in medium distance (1.5m-5m), and palm skeleton modeling when the target is near (0.5m-1.5m). These two modes can be switched automatically; 4. Instruction. In this article, the instruction is displaying the current action, it can be further extended.

In the actual recognition, the system works by three stages: firstly, scan the real-time frames, in order to confirm whether there are objects similar to skin color, and secondly when the probable target been found, the system confirms its authenticity, then establishes its skeleton model. At last, with the model's information, the target is tracked closely and the gestures are recognized automatically.

Image preprocessing

The process includes color segmentation, contour extraction and face recognition. The platform, which is used for real-time image processing, is built upon Opens.

Skin color segmentation. Skin color is an important physical characteristic, widely used in gesture and face recognition. The skin color can be utilized for image segmentation.

In computer vision, the commonly used color spaces are RGB, HSV, Crab, etc. According to the research by Brad ski G [9] which analyzes the distribution of different colors in different spaces and each color's performance in detection, it can be seen that in HSV, the skin color has better aggregation and constancy, mainly in the hue (H) selection. In this article, HSV is selected for color segmentation.

Through a lot of distributions of skin and non-skin colors in different scenes, the skin area is extracted according to the relations:

$$V = \begin{cases} 255 \ldots \text{if } H_{min} < V_{hue} < H_{max} \\ 0 \ldots \quad \text{otherwise} \end{cases}$$

Where V is a pixel in image I, and V_{hue} is its hue value. H_{min} And H_{max} are two parameters related to the distribution of skin in HSV.

A binary image where skin area is highlighted can be obtained after segmentation.

Contour extraction. After morphological operation on the binary image, looking for the contours with the function Find Contours, then the perimeter, area and some other information of each contour can be saved. In different areas of the human body, the ratio of perimeter and area of various contours is distinct. Thus face can be separated from hand precisely. Some interfere areas whether too narrow or too small are eliminated at the same time.

The size of the area can also reflect the distance between the target and the webcam [10-11].

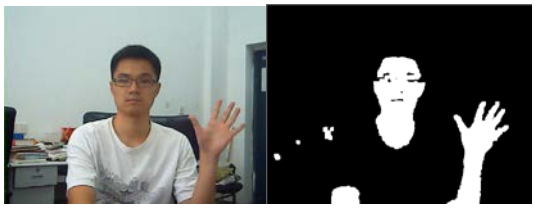

Fig.1 Skin segmentation

Fig. 2 Face detection

Face detection. In order to verify the detected face from contour extraction, Harry-Like is applied within the area's outline. If there is no true face found, then the system would scan the global frame.

Harry-Like is a calculation of the gray value between adjacent regions in an image [12-13]. It uses 14 rectangular prototype, which describe edge, linear and surrounded features. These features have certain discrimination to faces. After Adobos cascade, the successful classification rate is 98.7%.

Body modeling

In this article, the system expresses the gesture semantics with 2D models.

Body structure parameters. The human body is a very sophisticated system. It is a combination of various parts according to strict proportion and structure [14-15].

The proportion relationship used in the system is listed as follows:

Table 1 Body Structure Parameters

Head Height : Head Width	1:0.618
Shoulder Width : Head Width	3:1
Upper Arm Length : Forearm Length	3:2
Stature : Head Height	1.1:1
Palm Length : Palm Width	1:0.7

Joint positioning. At the initial time, when "begin" signed by the target is detected, the system calibrated the joints automatically.

Comparing to hands, human face's structure is more stable. Seek face contour's minimum circumscribed rectangle, then the system will regard the rectangle's length as the head's height, and the rectangle's width as the head's width.

Seek the centers in the face contour and hand contours, then the system regards the centers' positions as the head node and hand nodes' position.

Build the upper body skeleton and the joints according to the human structure. For body skeleton, the system marks the target's head, neck, spine, shoulders, elbows, hands and belly.

Build the palm skeleton and mark the joints with the function cvConvexHull2 and convexity Defects. For palm skeleton, the system marks its palm center, fingerprints, finger roots and wrist. The palm center is the hand contour's center. The fingerprints locating at the peaks of the convex hull, while the finger roots are at the valley of the convexity defects. Draw a circle at the palm center with a radius equaling the largest distance from the convex-hull point to the palm center. The wrist points are the intersection of the hand contour and the circle.

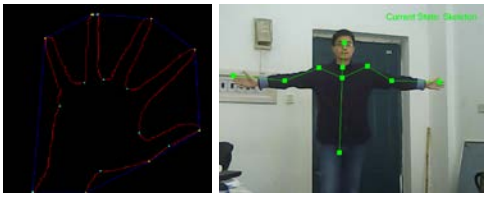

Fig.3 Hand Joints and upper body skeleton

Joints tracking. Referring to the Camshaft [16-17], the system:

1) Initial searching seeds at each joint,
2) Track these seeds,
3) Run Mean shift in every search box,
4) find the new center,
5) Update joint location to the correspond center,
6) set the seed location to the joint location,
7) Turn back to the 2nd step.

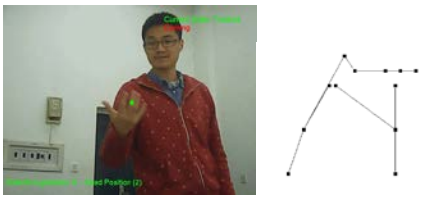

Fig.4 Trajectory of right hand in per second

Gesture recognition

Gestures compose of postures and movements. They can be recognized by locating every joint [18-20].

Posture Recognition. With regard to the feedback position of each joint, the system can acquire its coordinate $P_n(x,y)$, where n is the number of correspond joint. For each frame I, a vector $\overrightarrow{P}(P_1,P_2,\ P_3\ldots P_n)$ represent the posture is obtained. Classify $\overrightarrow{P}$ by GNC, the system can identify the current semantic.

Grouping neighbor classification (GNC). Based on KNN [21-22], GNC is a modified classification algorithm by optimizing the classification model and developing its self-learning ability.

Suppose there are N vectors, as for samples, divided into x groups through training, and there are c_x vectors in one group.

If there is an unknown vector, and the classification steps are:

a) Step.1: calculate the Euclidean distance $dist$ between the input vector and each center;

b) Step.2: take the minimum distance $mindist$ out, and save the group serial number I (0<i<x);

c) Step.3: calculate the closed distance $MinDist$ between every group center;

d) Step.4: if $mindist < \rho \cdot MinDist$, sort the vector into the ith group; otherwise make it the $(i+1)$th group; ρ is a parameter related to classification;

e) Step.5: update the center of the changed group;

f) Step.6: get the next unsorted vector;

g) Step.7: repeat the above steps.

This classification method can utilize a smaller storage, get a faster calculation and self-learning ability. It has a good stability for different postures in different situations.

Movement recognition. The changed position reflects the node's movement.

For one joint at $P_n(x,y)$, after 33ms, the system will get its new position $P'_n(x',y')$. Thus the joint's displacement and velocity can be obtained with $\Delta P_n(x'-x, y'-y)$. Classify these displacements and velocity, and the joint's movement can be recognized. The system can gain all direction, speed and path by comparing $\overrightarrow{P_i}$ in the ith frame and $\overrightarrow{P_{i+1}}$ in the $(i+1)$th frame.

Combine the posture and motion, the system can determine the actual significance of the current gesture.

For examples: recognize different angles between arms and the torso like the follow.

Fig.5 Sign"50°"

Fig.6 Raise right hand

Fig.7 Sign "5"

Conclusion

This paper introduces human body structure into the traditional monocular vision-based gesture recognition system. The optimized system builds skeleton model by locating the joints, and expresses complex gestures with a simple point-line model. It reduces the computation complexity and simplifies the recognition. This multi-feature fusion recognition enhances the accuracy and stability of gesture recognition.

In addition, the GNC is a self-learning algorithm. It is robust in different circumstances and classifies the gestures rapidly and precisely.

Acknowledgment

This study is financially supported by NSFC-CAS Joint Fund (No. U1332130), 111 Projects (No. B 07033), and 973 Project (No. 2014CB931804).

References

[1] Murthy G R S, Jadon R S. A review of vision based hand gestures recognition [J]. International Journal of Information Technology and Knowledge Management, 2009, 2(2): 405-410.

[2] Amayeh G, Bebis G, Erol A, et al. Hand-based verification and identification using palm–finger segmentation and fusion[J]. Computer Vision and Image Understanding, 2009, 113(4): 477-501.

[3] Lee D, Lee S G. Vision-based finger action recognition by angle detection and contour analysis [J]. ETRI Journal, 2011, 33(3): 415-422.

[4] Stergiopoulou E, Papamarkos N. Hand gesture recognition using a neural network shape fitting technique [J]. Engineering Applications of Artificial Intelligence, 2009, 22(8): 1141-1158.

[5] Rokade R, Doye D, Kokare M. Hand Gesture Recognition by Thinning Method [C] //Digital Image Processing, 2009 International Conference on. IEEE, 2009: 284-287.

[6] Feng Zhiquan, Jiang Yan. A Survey of hand gesture recognition [J]. Journal of University of Jinan (Sci. & Tech.), 2013, 4: 002.

[7] Alexiadis D S, Kelly P, Daras P, et al. Evaluating a dancer's performance using kinect-based skeleton tracking [C] //Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011: 659-662.

[8] Ren Z, Yuan J, Zhang Z. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera[C]//Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011: 1093-1096.
[9] Zarit B D, Super B J, Quek F K H. Comparison of five color models in skin pixel classification [C] // Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999. Proceedings. International Workshop on. IEEE, 1999: 58-63.
[10] Wahab M N A, Sivadev N, Sundaraj K. Target distance estimation using monocular vision system for mobile robot[C]//Open Systems (ICOS), 2011 IEEE Conference on. IEEE, 2011: 11-15.
[11] Han Y X, Zhang Z S, Dai M. Monocular vision system for distance measurement based on feature points [J]. Guangxue Jingmi Gongcheng (Optics and Precision Engineering), 2011, 19 (5): 1110-1117.
[12] Wang Y, Yuan B. A novel approach for human face detection from color images under complex background [J]. Pattern Recognition, 2001, 34(10): 1983-1992.
[13] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C] // Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2001, 1: I-511-I-518 vol. 1.
[14] Information on http://blog.sina.com.cn/s/blog_4b3730700100drg2.html
[15] Chen Guodong, Li Jianwei, Pan Lin, Yu Lun. Algorithm for extracting skeleton of 3D human body model based on body characteristic [J]. Computer Science, 2009, 36(7): 295-297.
[16] ZHANG H, ZHANG J, YUE H, et al. Object tracking algorithm based on CamShift [J]. Computer Engineering and design, 2006, 11: 031.
[17] Nouar O D, Ali G, Raphael C. Improved object tracking with CamShift algorithm [C] // Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. IEEE, 2006, 2: II-II.
[18] Yoon S M, Kuijper A. Human action recognition based on skeleton splitting [J]. Expert Systems with Applications, 2013, 40(17): 6848-6855.
[19] Giang N T, Tao N Q, Dung N D, et al. Skeleton based shape matching using reweighted random walks[C]//Information, Communications and Signal Processing (ICICS) 2013 9th International Conference on. IEEE, 2013: 1-5.
[20] Bian Z P, Chau L P, Magnenat Thalmann N. Fall detection based on skeleton extraction[C] // Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry. ACM, 2012: 91-94.
[21] Zhang H, Berg A C, Maire M, et al. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition [C] //Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. IEEE, 2006, 2: 2126-2136.
[22] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern recognition, 2007, 40(7): 2038-2048.