Interaction with 3D gesture and character input in Virtual Reality

Jiachen Yang, *Member, IEEE,* Yafang Wang, Zhihan Lv*, Na Jiang, and Anthony Steed

***Abstract*—Hand gesture recognition is a key aspect to make interaction of virtual reality more convenient. A good way to make users' idea understood by computers including characters input plays an important role in interaction. Current methods of hand gesture and character input are too limited to make full use of powerful capacity that computers have nowadays. In this paper, we propose a natural 3D input method based on stereo cameras as an interface of human and machine. We segment the hand out based on skin-color detection and train a neural network based on Hu moments to recognize valid and invalid gestures defined in our paper. For valid gestures, we implement stereo matching and 3D coordinate calculation and line them up to formulate characters. Our method can robustly recognize 3D gestures in different directions and make users' input more free compared with traditional ways.**

***Index Terms*—virtual reality, consumer electronics, three dimension interface, gesture recognition, binocular camera, character input.**

I. Introduction

WITH rapid development of consumer electronics, virtual reality (VR) is attracting great attention of research [1], [5], [6]. VR creates an immersive environment connecting the virtual world with real world [2]. Smooth interaction between human and computers is needed to make users feel what they see and hear is true. Input into various electronic equipment is a crucial part of interaction when users need to type a message into a virtual cellphone or control a game using gestures in a virtual environment. Nowadays most of the input in VR relies on wired equipment such as keyboards and handles, which confine users' activity to a limited range [4], [10], [11]. It cannot meet the demand of VR which should be based on convenient interaction. Thus much research has been done to find a more convenient way to input into equipment related to VR.

In an environment where human and computers are not connected directly, voice and body movements can be used to express our ideas in a good degree [7], [12], [14]. In terms of body movements, as the most flexible part of human's body, hand can be used most frequently to express us when we interact with computers in VR environment [15], [16].

Original ways of gesture recognition are usually based on some particular wearable device or color remark [10], [13], [37], [38], which make it possible to detect and track the movement of hands. However, these methods restrict the range of user's activities or the convenience of the experience. This became the bottleneck of human computer interface, which can be seen from the popularity of the gravity sensor-based gaming console. It engages the older generation who do not like to interact with electronic entertainment units using traditional keyboard and mouse. The development of WaveController also demonstrates that traditional computer and consumer electronics controls such as remote controllers can be replaced effectively with a hand gesture recognition system. Therefore, vision-based gesture recognition become the art-of-the-art trend [3]. Recognition of natural gestures enables that human can control whole VR system without wired equipment and make a free input into the system.

Traditional ways perform good with gestures in simple background, but they cannot deal with complex background and ignore the three-dimensional characteristics of hands. In this paper, we propose a natural 3D gesture recognition method used to record written characters in the air. It can be applied in the typing situations in VR environment to free the users out of various input devices. This will improve the efficiency and degree of freedom in the interaction between human and computers. By using stereo cameras, we can recognize 3D gestures with a high accuracy and more various angles compared with traditional 2D methods.

The rest of this paper is organized as follows. Section II introduces the related work briefly. Our system including hand model, gesture recognition and stereo calculation is presented in Section III. Section IV shows the experimental result and last section offers our view of conclusion.

II. Related work

Hand gesture recognition provides a natural, innovative and modern way of non-verbal communication. A detailed research about gesture recognition has been done in [9]. Most of the techniques rely on template matching [29] or shape descriptors [30]. A novel approach of hand gesture recognition based on detection of shape-based features is discussed in [17], but its performance can be easily influenced by constraints like hands' orientation. This method cannot be applied in real situations because of its inconvenience. To be more precise, hand is considered as a fully articulated object and a realtime hand tracking system using a depth sensor is presented in [8]. 3D model helps a lot in gesture recognition [18], [20]. Even though they are limited with high computational cost, the development from 2D to 3D recognition has moved human and

Jiachen Yang ang Yafang Wang are with School of Electronic Information Engineering, Tianjin University, Tianjin, P.R. China, 300072.

Zhihan Lv and Anthony Steed are with Department of Computer Science, University College London, London, UK, 33416327. E-mail: Z.Lu@cs.ucl.ac.uk. A.Steed@cs.ucl.ac.uk

Na Jiang is with Management services company of the first mining area, Da Gang Oilfield, PetroChina, Tianjin, P.R. China, 300072.

computer interaction a big step forward. Now with increase of computer capacity, neural networks can be a powerful tool to deal with adaptive gesture recognition [22]–[24].

Gesture recognition has a wide area of application including motion sensing game and VR [19]. A whole-hand input device called AcceleGlove can be used to manipulate three different virtual objects: a virtual hand, icons on a virtual desktop and a virtual keyboard using the 26 postures of the American Sign Language alphabet [26]. A system for navigating and acting in three-dimensional virtual environments by using hand gestures is presented in [25]. Fast response to gestures is required in hand controlled games [15].

In this paper, we proposed a natural 3D gesture recognition algorithm, which can be used to record users' written characters when users interact with computers. Characters can be divided into smaller stroke units. We define two gestures to indicate users are writing strokes or their hands are just moving between two strokes. Through a trained neural network based on Hu moments, we can check which kind of gesture it is in front of the camera. Stereo matching and depth calculation is then carried out to determine the hand's location in real word. These locations are recorded and connected to formulate characters.

III. Hand Modeling and Recognition

A. Hand segment

We use binocular camera to catch users' gestures. The first thing to comprehend these gestures is to find out where the hand is and cut it out of the background. Color detection is one of the most common methods to determine rough location of hands in images.

Skin color has different distribution in different color spaces, so a good color space can make great contribution to the accuracy of hand detection. We choose $YCbCr$ because skin colors gather in an ellipse region which is easy to detect. In $YCbCr$ space, luminance and chromaticity are independent, and Y indicates luminance of colors and Cb, Cr corresponds to the chromaticity of blue and red separately. It reduces the redundancy present in RGB color channels and represents the color with statistically independent components. $YCbCr$ is one of the most popular choices for skin detection [34].

Images catched by cameras are in RGB space, but they can be transformed into $YCbCr$ through formula (1), which is pixel-based calculation.

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

Since the mapping is linear with RGB information, luminance Y cannot be totally independent on chromaticity. It results in that skin region change nonlinearly with luminance. To avoid the influence of luminance, a nonlinear transform of $YCbCr$ is carried out and then we get an ellipse region where skin color gathers. The ellipse model are given in formula (2) and (3), where $a = 25.39$, $b = 14.03$, $ec_x = 1.60$, $ec_y = 2.41$,$c_x = 109.38$, $c_y = 152.05$.

$$\frac{(x - ec_x)^2}{a^2} + \frac{(y - ec_y)^2}{b^2} = 1 \quad (2)$$

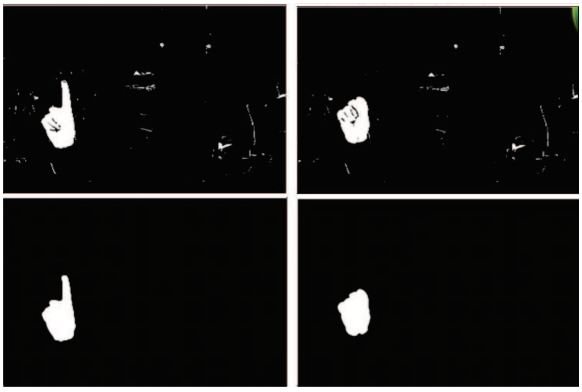

Fig. 1. Hand segment and feature extraction. First row: after skin color detection. Second row : after morphological process.

Fig. 2. Morphological process.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} Cb - c_x \\ Cr - c_y \end{bmatrix} \quad (3)$$

A pixel belongs to a hand or not is up to that whether its color locates in the ellipse region. If it is, the pixel is set to be 1 and otherwise 0 such that we can roughly detect the hand in images.

However, it should be noticed that there might be something in the background whose color is similar with skin. Then noise is introduced into the hand region detected by color detection (seen in Fig. 1). Under the premise that there is no large skin-colored object other than hands in images, we choose the region with largest area as the target hand region. Noise unconnected with hand region can be removed then. In terms of the noise connected with hand, morphological process can be used to remove it (seen in Fig. 2). Firstly erosion and dilation is carried out to make the region smooth and then filtering to remove disturbing noise. At last holes in the region are filled to make a complete hand. Till now, we get binary images with hand regions are labeled with 1.

B. Gesture recognition

Chinese characters are combination of strokes. When users write characters, their hands move in some particular order. Some move corresponds to strokes and some is just transition between strokes. We define two gestures to represent these two states of hand movements (seen in Fig. 3). The valid gesture, a hand with index finger raised up, indicates users are writing strokes. The invalid gesture, a closed hand, indicates users' hands are moving between strokes.

1) Feature extraction: There are so many attempts to extract various features from hands for gesture classification [35], [36]. The feature extraction must capture the essence of a gesture. It is very critical in hand gesture recognition

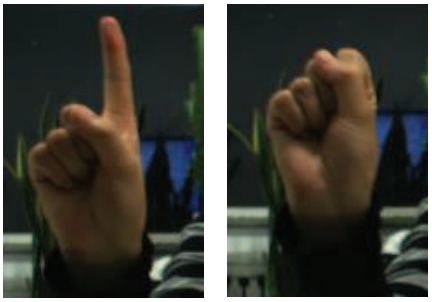

Fig. 3. Gesture definition. (a) Valid gesture used when users write strokes. (b) Invalid gesture used when users' hands move between strokes.

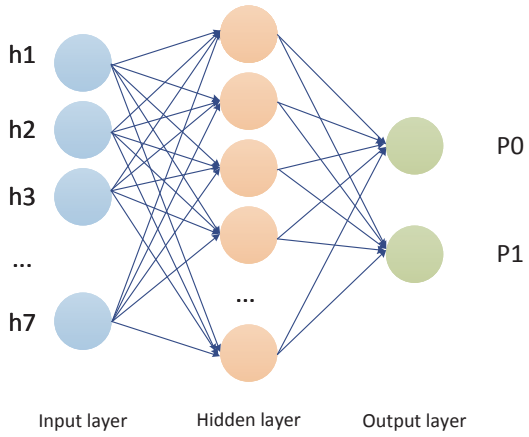

Fig. 4. Feed-forward network with a single hidden layer.

because hand is non-rigid. Gesture variations caused by rotation, scaling and translation can be circumvented by using a set of features that are invariant to these operations, such as moment invariants. Hu moments have been considered as one of the most effective methods to describe deformable object and they are widely applied in classification of subjects [27]. Essentially, Hu moment algorithm derives a number of self-characteristic properties from a binary image of an object. These properties are invariant to rotation, scale and translation [21].

For a digital image $f(x, y)$, with size M*N, moment $m_{p,q}$ is defined as formula (4) and central moment $\mu_{p,q}$ is defined as formula (5), where $\bar{x}$ and $\bar{y}$ are the coordinates of image center.

$$m_{p,q} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) x^p y^q \tag{4}$$

$$\mu_{p,q} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y)(x - \bar{x})^p (y - \bar{y})^q \tag{5}$$

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}} \tag{6}$$

Then we get normalized moments as

$$\eta_{p,q} = \frac{\mu_{p,q}}{m_{00}^{\frac{p+q}{2}+1}} \tag{7}$$

Hu moments are linear combinations of the central moments and seven moments are given as follows.

$$\begin{aligned}
h1 &= \eta_{20} + \eta_{02} \\
h2 &= (\eta_{20} - \eta_{02})^2 + 4{\eta_{11}}^2 \\
h3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
h4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
h5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 \\
&\quad -3(\eta_{21} + \eta_{03})^2) + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\
&\quad (3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
h6 &= (\eta_{20} - \eta_{02})((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
&\quad +4{\eta_{11}}^2(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
h7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 \\
&\quad -3(\eta_{21} + \eta_{03})^2) + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \\
&\quad (3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2)
\end{aligned} \tag{8}$$

These features effectively describe hand gestures and we use them in our system for classification using neural network.

2) Network learning and training: Neural network is powerful to model non-linear relationship. It has been applied to perform complex functions including pattern recognition [31], classification [32], identification [33] and so on. It is also able to learn and predict over the time just as a human-like entity.

We exploit a feed-forward network with a single hidden layer (seen in Fig. 4). It is trained using back-propagation in which Hu moments and corresponding gesture labels are used as input and output vectors to train the network until it approximates a function between input and output [28]. There are only three layers due to the limited number of hand gestures to be recognized. Since there are seven Hu moments and two gestures defined, our neural network model has seven input neurons and two output neurons, and there are 5 neurons in the hidden layer.

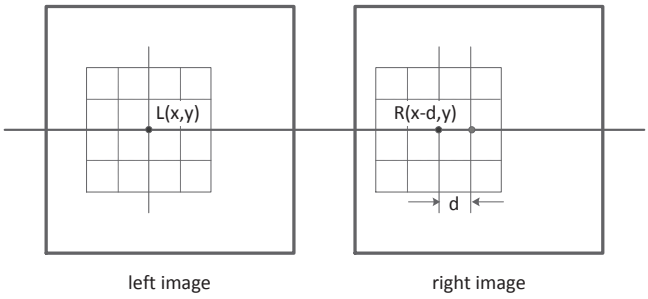

Fig. 5. Stereo matching based on region similarity. $L(x, y)$ is the point to be matched and $R(x - d, y)$ is its corresponding point in right image.

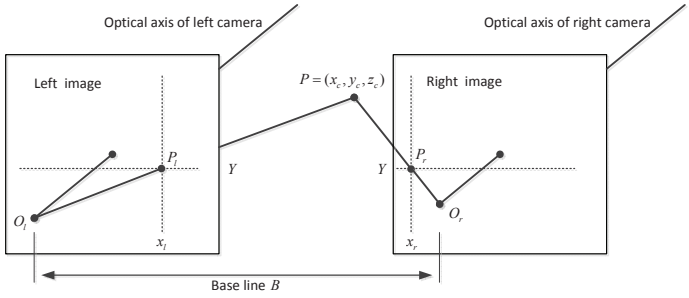

Fig. 6. Map of locations of real world system and image system. O_l and O_r are optic center of camera lenses. P is the coordinate in real world, corresponding to P_l and P_r in left and right image

C. Stereo calculation

We detect hand in the sequent images based on color detection and segment it out. After hand segment, we get the location of the hands. To reduce calculation cost, we check the location for every m frames because movement change is small in adjacent frames. The gap m is an experience value.

Detection are carried out on one of image pair catched by binocular cameras. We analyze the gestures before their 3D location in real world because only location of valid gestures are concerned in our paper. Once a valid gesture is found, following will be stereo matching and coordinates calculation of the mass center of the hand.

Stereo matching is to find corresponding point in a pair of image catched by binocular cameras. As the difference of cameras' shooting location, one point has different locations in right and left image. We implement stereo matching based on region similarity (seen in Fig. 5). To find the corresponding point of $L(x, y)$ in left image, a small window centered on $L(x, y)$ is determined and its gray-scale histogram is calculated as its feature. Search in right image with the window to find the region with largest similarity based on histogram information.

The premise of realizing stereo matching for image pair is that two cameras are in same horizontal level and their optic axes are parallel. Then an object can be targeted in real world based on the matching point and its locations in left and right image. For a point $P(x_c, y_c, z_c)$ in real world, it corresponds

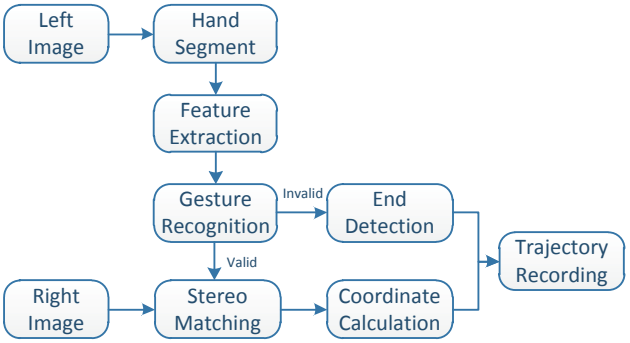

Fig. 7. Framework of whole algorithm.

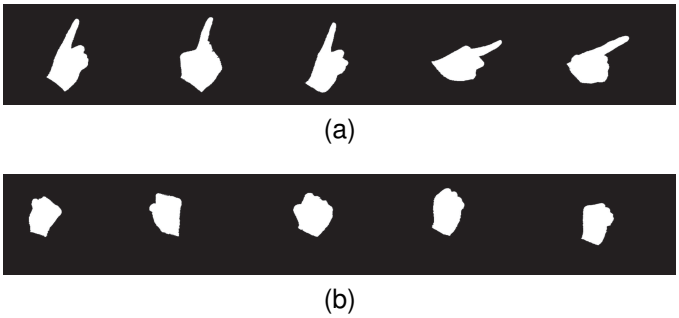

Fig. 8. Training samples. (a)valid geture samples. (b)invalid geture samples.

to $P_l(x_l, y_l)$ in left image and $P_r(x_r, y_r)$ in right image (seen in Fig. 6). They follows the relationship in formula (9), where B is the distance between the optical centers of two cameras and f indicates their identical focal length.

$$\begin{cases} x_c = \frac{B \cdot x_l}{x_l - x_r} \\ y_c = \frac{B \cdot y_l}{x_l - x_r} \\ z_c = \frac{B \cdot f}{x_l - x_r} \end{cases} \tag{9}$$

D. Character recording

We define two gestures in Section B and valid ones are key to character trajectory, so only valid gestures are processed with stereo matching and coordinate calculation. If a valid gesture is found, we need to judge wether the strokes come to an end. We consider continuous detection of invalid gestures as a sign of strokes' end. Then all calculated coordinates in a stroke are connected to formulate the written characters. By mapping them into a normalized plane, these characters can be recognized by OCR technique offered by Hanwang software which is useful in Chinese character recognition. Till now, characters written by users with natural hand are input into computer successfully. The framework of whole algorithm is in Fig. 7.

IV. Experiment and Evaluation

We take 200 gesture samples to train the neural network, including 100 valid gestures and 100 invalid gestures. To

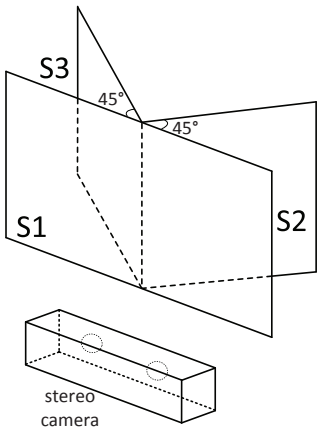

Fig. 9. Writing planes in the experiment. Plane S1 is parallel to imaging plane of stereo camera. S2 rotates to left and has a angle of 45 degree compared with S1. S3 rotates to right and also has a angle of 45 degree compared with S1.

TABLE I
ACCURACY OF GESTURE RECOGNITION. NUM(CHECKED) MEANS THE NUMBER OF ALL GESTURES THAT ARE CHECKED IN OUR EXPERIMENT AND NUM(RIGHT-LABELED) MEANS THE NUMBER OF RIGHT RECOGNIZED GESTURES, INCLUDING VALID AND INVALID GESTURES. ACCURACY IS THE PERCENTAGE OF RIGHT-LABELED GESTURES COMPARED WITH THE NUMBER OF CHECKED GESTURES.

Planes	$S1$	$S2$	$S3$	*sum*
$num(checked)$	1232	1084	1165	3481
$num(right-labeled)$	1162	1013	1085	3228
recognition accuracy	0.94	0.93	0.93	0.936

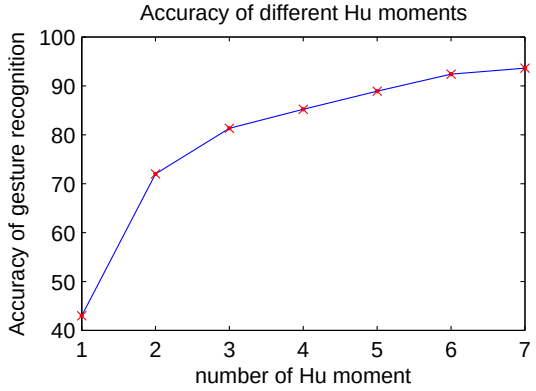

Fig. 10. Accuracy of gesture recognition depend on different numbers of Hu moments.

increase the flexibility of the network, gestures are sampled in different directions (seen in Fig. 8). The camera in the experiment is RoHS 1024*768 Color 3.8 mm Bumblebee2 camera. To evaluate the performance of our system, we write 3 groups of characters in the air using the gestures we have defined. Each group of characters are the same character written in three different planes (seen in Fig. 9). Plane S1 is parallel to imaging plane of stereo camera. S2 rotates to left and has a angle of 45 degree compared with S1. S3 rotates to right and also has a angle of 45 degree compared with S1,

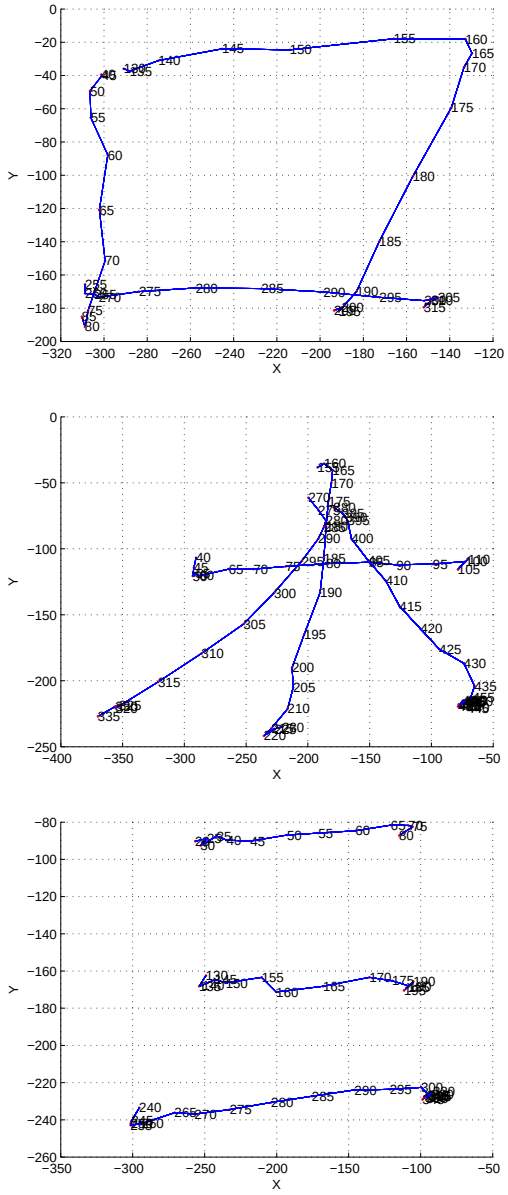

Fig. 11. Recording trajectory of writing characters directly facing the stereo camera, corresponding to $S1$.

just as seen in Fig. 9. We choose these different writing planes to test our system's flexibility to deal with various free input which means there is no restriction that users must write the character strictly facing to the camera. Tracking gap m in our experiment is 5.

A. Gesture recognition

In our experiment, every gesture checked is labeled as valid or invalid by the trained network. Firstly to evaluate the performance of Hu moments, we carry out an experiment to find out different number of Hu moments' effect on the gesture recognition. As seen in Fig. 10, the number of hu moments represents how many features we used from h1 on, eg. number 1 means we use only h1, number 3 means we use h1, h2 and h3. We can see that the accuracy of gesture

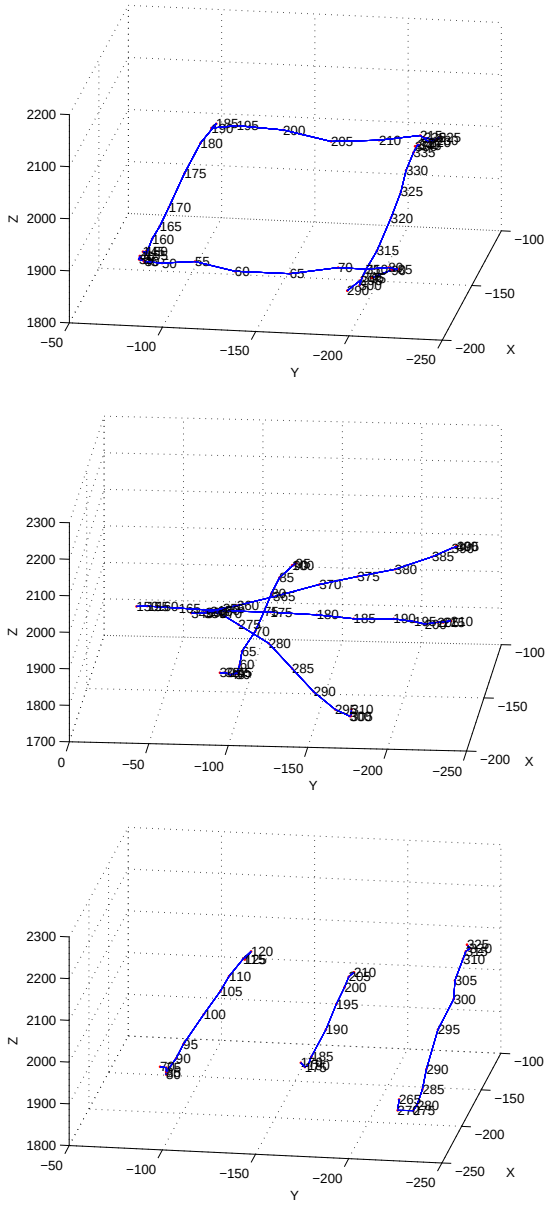

Fig. 12. Recording trajectory of writing characters facing to the left of the stereo camera, corresponding to $S2$.

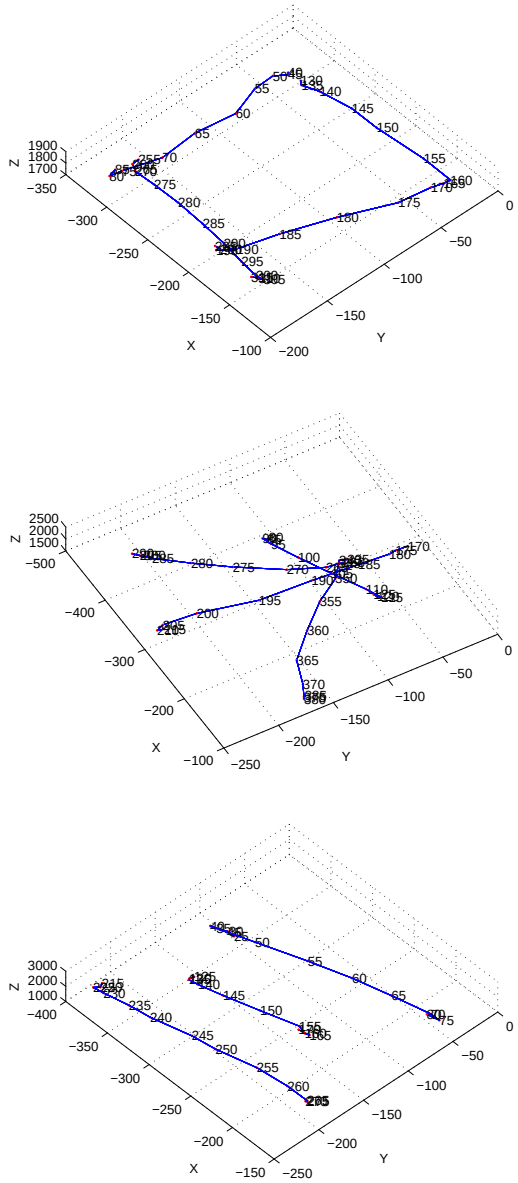

Fig. 13. Recording trajectory of writing characters facing to the right of the stereo camera, corresponding to $S3$.

recognition increases with the number of features used, and the increasing speed is more rapid when the number is small.

To evaluate the performance of our gesture recognition, we count the total number of gestures that are checked and the number of right-labeled ones. As seen in Table I, we make a statistics in three writing planes. We can see that the accuracy in three different direction are equally high. One reason is that Hu moment describe the gestures in a robust way. On the other hand, the strategy of sample selection also increase the network's ability to adjust to shape change. This shows that our gesture recognition is able to deal with free input in different directions and this will greatly increase the convenience when users interact with computers in VR environment.

B. Character recording

With stereo cameras as input device, 3D location of the targets can be obtained conveniently. Combined with the writing gestures we have defined, we can obtain the three dimensional trajectory of users' hand. For a valid gesture, we calculate its coordinates in real world using the technics explained in Section III.D and connect them to formulate the written characters. Fig. 11, Fig. 12 and Fig. 13 show some recording trajectory of our input characters. As each sample character is written in three different planes, all characters are recorded three times corresponding to different planes. We can see that the characters are recorded correctly regardless of the writing planes. This demonstrates that our algorithm can successfully record 3D characters. After these characters

are normalized, these characters are recognized correctly by Hanwang OCR v8 [39]. This shows that free 3D input is successfully realized.

We apply this character input method in VR environment to free users out of the limit of various devices. As it can realize natural 3D hand gestures' recognition, it will be more convenient and more accurate because of our feature selection and network training work. For the character recording, stereo calculation makes it more robust than traditional 2D methods.

Accurate trajectory recording and high recognition accuracy indicate high communicating capacity of this gesture interacting method with computers. When controlling the targeted VR devices or texting massages into VR system, 3D natural gesture recognition and character recording can offer more information than traditional ways. However, the 3D information used in the algorithms and the training process can be more complex which would make the algorithms a bit slower than traditional ones.

V. Conclusion

In this paper, we propose a natural 3D input method based on stereo camera as an interface of human and machine. We segment the hand out based on skin-color detection. To distinguish the valid input and invalid input in characters, two gestures are defined in our method. Taking advantage of Hu moments' invariance and neural network's intelligence, we can recognize these two gestures with high accuracy. For valid gestures, we implement stereo matching and 3D coordinates calculation and line them up to formulate characters which can be recognized by OCR technique. This method makes touch-less interaction in virtual reality possible and its robustness enables users' free input.

Acknowledgment

This research is partially supported by the National Natural Science Foundation of China (No.61471260 and No.61271324) and Natural Science Foundation of Tianjin (16JCYBJC16000).

References

[1] Koverman C. Next-Generation Connected Support in the Age of IoT: It's time to get proactive about customer support[J]. IEEE Consumer Electronics Magazine, 2016, 5(1): 69-73.

[2] Markwalter B. Entertainment and Immersive Content: What's in store for your viewing pleasure[J]. IEEE Consumer Electronics Magazine, 2015, 4(1): 83-86.

[3] Brun L, Gasparini A. Enabling 360 visual communications: next-level applications and connections[J]. IEEE Consumer Electronics Magazine, 2016, 5(2): 38-43.

[4] Yang J, Xu R, Lv Z, et al. Analysis of Camera Arrays Applicable to the Internet of Things[J]. Sensors, 2016, 16(3): 421.

[5] Yang J, Ding Z, Guo F, et al. Multiview image rectification algorithm for parallel camera arrays[J]. Journal of Electronic Imaging, 2014, 23(3): 033001-033001.

[6] Tzionas D, Gall J. 3D Object Reconstruction from Hand-Object Interactions[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 729-737.

[7] Bambach S, Lee S, Crandall D J, et al. Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 1949-1957.

[8] Qian C, Sun X, Wei Y, et al. Realtime and robust hand tracking from depth[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1106-1113.

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

[10] Ishiyama H and Kurabayashi S, Monochrome glove: A robust real-time hand gesture recognition method by using a fabric glove with design of structured markers[C]. Virtual Reality (VR), 2016 IEEE. IEEE, 2016: 187-188.

[11] Guo S, Zhang M, Pan Z, et al, Gesture Recognition Based on Pixel Classification and Contour Extraction[C]. 2015 International Conference on Virtual Reality and Visualization (ICVRV). IEEE, 2015: 93-100.

[12] Vafadar M, Behrad A. A vision based system for communicating in virtual reality environments by recognizing human hand gestures[J]. Multimedia Tools and Applications, 2015, 74(18): 7515-7535.

[13] Lv Z, Feng S, Feng L, et al. Extending touch-less interaction on vision based wearable device[C]. 2015 IEEE Virtual Reality (VR). IEEE, 2015: 231-232.

[14] Rosa-Pujazón A, Barbancho I, Tardón L J, et al. Fast-gesture recognition and classification using Kinect: an application for a virtual reality drumkit[J]. Multimedia Tools and Applications, 2015: 1-28.

[15] Wachs J P, Kölsch M, Stern H, et al. Vision-based hand-gesture applications[J]. Communications of the ACM, 2011, 54(2): 60-71.

[16] Liu Y, Yin Y, Zhang S. Hand gesture recognition based on HU moments in interaction of virtual reality[C].Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on. IEEE, 2012, 1: 145-148.

[17] Panwar M, Mehra P S. Hand gesture recognition for human computer interaction[C]. Image Information Processing (ICIIP), 2011 International Conference on. IEEE, 2011: 1-7.

[18] Guan H, Feris R S, Turk M. The isometric self-organizing map for 3d hand pose estimation[C]. 7th International Conference on Automatic Face and Gesture Recognition (FGR06). IEEE, 2006: 263-268.

[19] Cui Y, Weng J. Appearance-based hand sign recognition from intensity image sequences[J]. Computer Vision and Image Understanding, 2000, 78(2): 157-176.

[20] Ming Y. Hand fine-motion recognition based on 3D Mesh MoSIFT feature descriptor[J]. Neurocomputing, 2015, 151: 574-582.

[21] Hu M K. Visual pattern recognition by moment invariants[J]. IRE transactions on information theory, 1962, 8(2): 179-187.

[22] Xu D. A neural network approach for hand gesture recognition in virtual reality driving training system of SPG[C]. 18th International Conference on Pattern Recognition. IEEE, 2006, 3: 519-522.

[23] Huang Y S, Wang Y J. A Neural-Network-Based Hand Posture Recognition Method[M]. Transactions on Engineering Technologies. Springer Netherlands, 2014: 187-201.

[24] Molchanov P, Yang X, Gupta S, et al. Online Detection and Classification of Dynamic Hand Gestures With Recurrent 3D Convolutional Neural Network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4207-4215.

[25] Maggioni C. A novel gestural input device for virtual reality[C]. Virtual Reality Annual International Symposium, 1993, 1993 IEEE. IEEE, 1993: 118-124.

[26] Hernandez-Rebollar J L, Kyriakopoulos N, Lindeman R W. The AcceleGlove: a whole-hand input device for virtual reality[C]. ACM SIGGRAPH 2002 conference abstracts and applications. ACM, 2002: 259-259.

[27] Premaratne P. ISAR ship classification; An alternative approach[J]. CSSIP-DSTO Internal Publication, 2003.

[28] Huang D. Radial basis probabilistic neural networks: model and application[J]. International Journal of Pattern Recognition and Artificial Intelligence, 1999, 13(07): 1083-1101.

[29] Shan C, Wei Y, Qiu X, et al. Gesture recognition using temporal template based trajectories[C]. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004, 3: 954-957.

[30] Harding P R G, Ellis T. Recognizing hand gesture using Fourier descriptors[C]. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004, 3: 286-289.

[31] Chen W S, Yuen P C, Huang J, et al. Kernel machine-based one-parameter regularized fisher discriminant method for face recognition[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2005, 35(4): 659-669.

[32] Căleanu C, Huang D S, Gui V, et al. Interest Operator versus Gabor filtering for facial imagery classification[J]. Pattern recognition letters, 2007, 28(8): 950-956.

[33] Zhao Z Q, Huang D S, Sun B Y. Human face recognition based on multi-features using neural networks committee[J]. Pattern Recognition Letters, 2004, 25(12): 1351-1358.
[34] Kakumanu P, Makrogiannis S, Bourbakis N. A survey of skin-color modeling and detection methods[J]. Pattern recognition, 2007, 40(3): 1106-1122.
[35] Chen Q, Georganas N D, Petriu E M. Real-time vision-based hand gesture recognition using haar-like features[C]. 2007 IEEE Instrumentation Measurement Technology Conference IMTC 2007. IEEE, 2007: 1-6.
[36] Khamis S, Taylor J, Shotton J, et al. Learning an efficient model of hand shape variation from depth images[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2540-2548.
[37] Lv, Z., Halawani, A., Feng, S., Li, H., Rhman, S. U. (2014). Multimodal hand and foot gesture interaction for handheld devices. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 11(1s), 10.
[38] Lv, Z., Halawani, A., Feng, S., Ur Rhman, S., Li, H. (2015). Touchless interactive augmented reality game on vision-based wearable device. Personal and Ubiquitous Computing, 19(3-4), 551-567.
[39] http://ka.hanwang.com.cn/en/

Jiachen Yang received the M.S. and Ph.D. degrees in Communication and Information Engineering from the Tianjin University, Tianjin, China, in 2005 and 2009, respectively. He is an associate professor at Tianjin University. He was also a visiting scholar in the Department of Computer Science, School of Science at Loughborough University, UK. His research interests include stereo camera, stereo vision research, pattern recognition, stereo image displaying and quality evaluation.

Yafang Wang is a M.S. student with school of Electronic Information Engineering, Tianjin University, Tianjin, China. Her research interests include intelligent transportation systems, machine learning and stereo vision research.

Zhihan Lv was granted PhD. degree in Computer applied technology from Ocean university of China. Since 2012, He have held an assistant professor position at Chinese Academy of Science. His research interests include virtual reality, augmented reality, multimedia, computer vision and human-computer interaction.

Na Jiang Graduated from Communication and Information Engineering School In Tianjin University of Technology, Tianjin, China. Now she is working at Management services company of the first mining area, Da Gang Oilfield, PetroChina.

Anthony Steed is a Professor in the Virtual Environments and Computer Graphics group in the Department of Computer Science, University College London. He is currently head of the group. His research area is real-time interactive virtual environments, with particular interest in mixed-reality systems, large-scale models and collaboration between immersive facilities.