Research on a Visual Sensing and Tracking System for Distance Education

https://doi.org/10.3991/ijet.v14i08.9827

Fei Yang, Rong Zhang (✉), Youpeng Zhao
Wuhan University, Hubei, China
zhangrzrong@126.com

Abstract—With the Microsoft's RGB-D sensor Kinect and a customized 3-axis pan-tilt-roll machinery, a new intelligent sensing and tracking system was designed and manufactured. In order to simulate human natural visual sensing behavior, this system adopts the sensing function from Kinect installed upon the 3-axis motion machinery, controlled by an expert PID control algorithm based on the adaptive Kalman filter, so as to guarantee automatic real-time visual tracking and to observe human movements and receive his/her position information. This system is designed to be applied directly to the video production of distance education, aiming to improve distance education itself and resolve difficulties in keeping account of users learning state. It also has great potential of functioning as a basic platform where many other human-computer natural interactions can be extended.

Keywords—Distance Education; human-computer natural interactions; visual sensing; automatic tracking

1 Introduction

With rapid developments of information technology and increasingly popularity of multimedia technology, the time of modern distance education such as MOOC (Massive Open Online Courses) and micro learning has come. What differentiates it from traditional education most is that it does not submit to the limitations of location and time, displaying huge advantages over traditional education: more courses, genres and learning resources accessible on the Internet? However, the defect of lacking human-computer interactions stands in the way: users have great difficulties in finding these online courses as intriguing as traditional ones, during which they could interact with lecturers on scene, reading their signature body language, mainly due to the fact that only three versions of course materials as video, audio and text can be presented to the audience. The desire to learn would hardly be aroused which make users failed to enjoy staring at the screen, eventually dramatically increasing the possibility of falling into abeyance and damaging users' mental health. The technique of building a natural human-computer interaction system for users, thereby, is a must-solve issue for distance education [1].

The reigning popular sensing technology might provide a viable solution. Sensing technology stands out as one technology where a device that detects the motions of nearby persons or objects through sensory receptors, then respond right back after automatic analysis[2]. The core idea of promoting the performance for computers to sense human intentions and replicating the daily human-human interactions, has diminished human's dependence of input devices such as mouses, keyboards and remote controllers. Though being the most natural interaction technique alive, sensing devices actually rely largely on visual receptions, which make up more than 70% of human receptions. Undoubtedly, visual sensing system has substantial significance in human-computer interaction system. For present, few visual interaction systems arrive at the market designed for distance education, and most of which are, however, only for gaming experience and other special fields.

In this paper, a method of building a new intelligent sensing and tracking system, with Kinect and a customized 3-axis pan-tilt-roll machinery combined together is presented. This specified system receives position information of the lecturer through Kinect placed upon the three-axis motion machinery, controlling the camera in Kinect (or any other digital cameras) to track the movements of the lecturer so as to capture his/her motions and facial expressions to present to the audience, with no more professional photographers needed to frequently adjust positions to shoot the film and largely decreasing the workload for distance education. Changeable visual angles greatly improve the realism for audience, and fully present the poses, facial expressions along with untouchable audio, making distance courses much more immersive and interesting, improving distance education itself and resolve difficulties in keeping account of users' learning state. Besides, the real-time detection of the lecture's movements gives the lecturer him/herself no worries to clumsily control all kinds of sophisticated teaching techniques and allows for more focus on preparing courses, which shares no distinction with traditional ones. Applied to distance education, or any live lectures, commercial speeches and presentations, this system can surely play its part and make a difference. Moreover, the human-computer interaction based on this platform could be more promising in the future.

2 Selection of Visual Sensors

In recent years, it has given rise to many visual motion sensors designed for ordinary consumers. With relatively affordable RGB-D sensors, not only can users easily receive the multicolored RGB information, but also the exact 3-dimension position of each part in human bodies and motion data. Three products emerging in the market, Prime Sense Carmine, Microsoft Kinect and Asus Xtion, all of which adopted the light coding technique from the Israel Prime Sense Corporation, which belongs to one of structured light techniques. Different from traditional methods, the laser speckle of light coding possesses high randomicity, varying with distance. For example, the structured light reflects marks every position in the field, if one object was put into the field, and the exact position of the object could be acquired once the laser speckle of the object's surface is obtained. Also, ordinary CMOS chips can achieve this goal, which greatly cut down the expenses of manufacturing. Microsoft got technology from Prime Sense to produce Kinect, as the input device for Xbox 360 into the market in 2010. Kinect was equipped with three cameras, the middle one as multicolored RGB camera, the other two as the infrared laser projector and infrared CMOS camera. Players can use Kinect solely to play video games with no more controllers. Moreover, Microsoft opened up ports for all sorts of resources for personal computers, which creates numerous opportunities for developers to do whatever they like. Kinect has become not only a video gaming input device, but also a major tool for industrial, commercial, educational and medical industry to count on in the future[3]. Afterwards, ASUS's Xiton came out, which adopted the same design from Prime Sense, sharing much similarity with Carmine from Prime Sense itself. After Kinect being fashionable around the world, Microsoft quickly started its move to develop the next generation for Xbox One - Kinect One, which is of higher resolution.

Apart from the premiere products, companies like Intel, Creative, Leap Motion and Code Laboratories also launched their own visual sensing devices accordingly. Intel developed RealSense Series, designed for short-distance recognition of human gestures and expressions; RealSense Series are mostly integrated into laptop screens for sale; F200 is another series produced by Creative and Intel, with which consumers can easily get their hands on. Besides, there was a sensing camera called Senz3D that Intel launched together with Creative in 2013, which focused on gesture recognition and facial recognition, using the depth sensing sensors developed by Soft Kinetic. During the same period, Leap Motion was brought to the world by Leap Motion, focusing on gesture control. F200, Senz3D and Leap Motion share the similarity that they are all designed for short-distance application such as users sitting in front of the computer, and has little capability to detect the whole body. Some differences between the above sensors are shown in Table 1.

Table 1. Comparisons between sensors

Device	Xtion	Kinect (Xbox One)	F200	Senz3D	Leap Motion
Detection Range	Whole Body	Whole body	Upper body	Upper body	Front elbow
Difficulty for Developing	Relatively easy	Easy	Relatively easy	Relatively easy	Relatively easy
Working Distance	middle	middle	short	short	short

In conventional classes, an appropriate distance between students and teachers not only guarantees a broader vision, but also assuages pressure and anxiety for both sides, which Kinect can perform perfectly. Compared with other sensors, Kinect is easy to buy in the market and is with more tools and documents facilitate the development, so we adopted it as a fundamental sensor for the sensing and tracking system. Because any devices presented above can only be applied for a limited field of view, the customized 3-axis pan-tilt-roll machinery widening the field of view should be designed to meet the needs for practical applications.

3 Automatic Detection of Body

It always be a hot topic to automatically detect the position of human from static images and video sequences, with algorithms as the background-difference, neighboring frame subtraction, optical flow, particle filter [4] and HOG-based statistic method. For 2D image, due to the varieties of movements and clothing, along with lighting conditions and complex backgrounds, algorithms above do not show much promising results [5] . However, Kinect, turns things around. On one hand, detecting human body from depth images by active infrared structured light instead of traditional 2D colored images makes it possible to work under unstable lighting condition, even in complete darkness [6]. On the other hand, the algorithm based on depth images proposed by Shotton et al. proves to be an effective machine learning algorithm. It firstly analyzes millions of depth images, and categorizes them into random decision forests to rapidly recognize each part of human skeleton, and then uses pruning optimization by Bayesian algorithm to improve the precision and speed for recognition [7]. In this paper, adopting this algorithm, we can automatically detect and identify each part of body and thus infer 20 feature joints (including three types: traceable, untraceable and presumed), with which user can assess each part's information received from different feedback. In experiments, we use a computer equipped with Intel i7-3770 CPU, producing 30 results per second. An example of skeleton with 20 automatically detected feature joints is shown in Fig. 1.

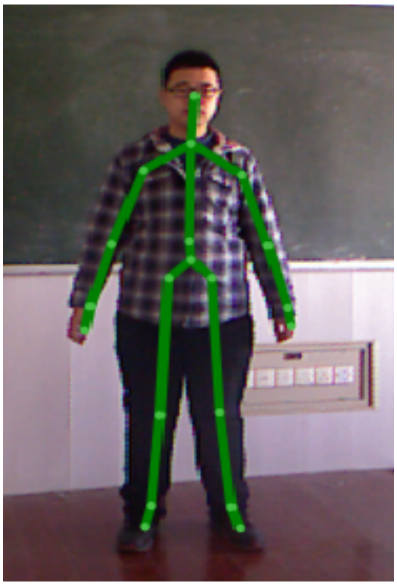

Fig. 1. The detected skeleton and 20 feature joints

From multiple experiments, we can conclude that the 20 feature joints can be easily recognized accurately when the body faces Kinect (see Fig. 1). However, when looking at the flank or blocking part of body, not all feature joints can be located, and some of which requires algorithm to further calculate. In Fig. 2, detection results of common positions are shown. The dots connected by thin lines are presumed joints, and the dots connected by thick lines are joints with. From a great of experiments, we can also conclude that feature joints between two shoulders have better positioning accuracy, which is not influenced by the position, distance, dressing, and other factors in various teaching scenarios. Moreover, the joint is located right below the head, the position where people focus most (the facial part) when interacting, sharing the approximately same height of blackboard. Therefore, in this paper, this joint is selected as target joint of tracking, with Kinect following its movement, to stimulate the natural interacting behaviors of students.

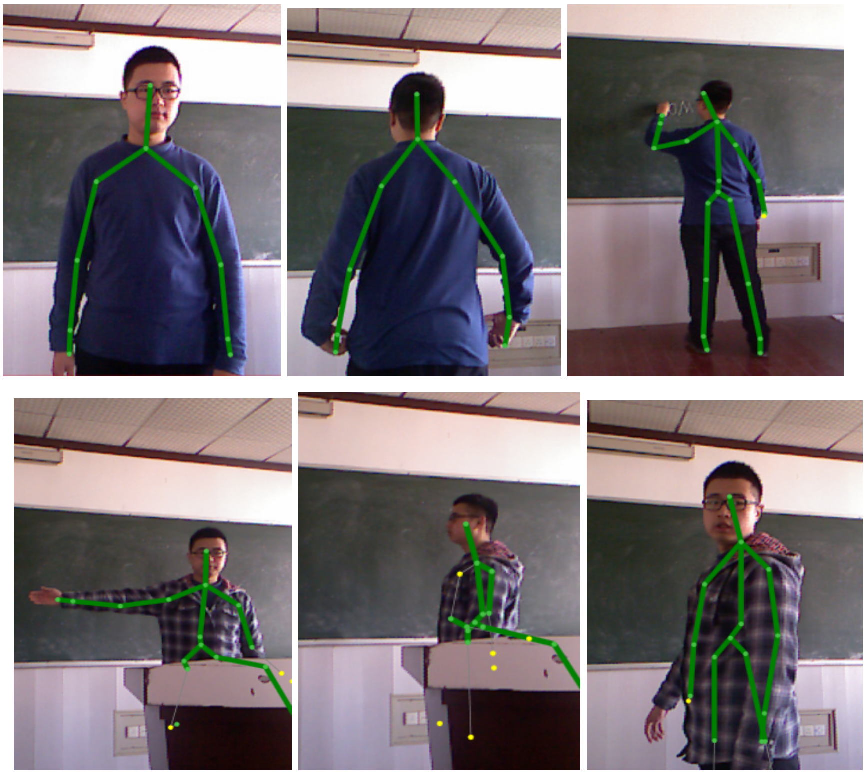

Fig. 2. Some experiment results in teaching scenarios

4 Auto-Tracking Platform and its Control

4.1 Hardware design of auto-tracking platform

A 3-axis pan-tilt-roll machinery is designed to realize the real-time posture control of Kinect (or any other cameras) for the purpose of simulating active adjustments of heads and expanding the sensing range. As shown in Figs. 3 and 4, the mechanical design and the manufactured device are presented. The 3-axis pan-tilt-roll machinery consists of three independent axes of rotation and servomechanisms, and each one can be separately controlled by its own motors. The maximum rotation range of yaw axis, pitch axis and transverse roller is 360°, 180°and 360°respectively. To achieve high control accuracy, three gyroscopes are fixed on each axis to measure its angular speed and send feedback to the controller. The controller communicates with the computer by wireless serial port, and receives orders from the algorithm program in the computer.

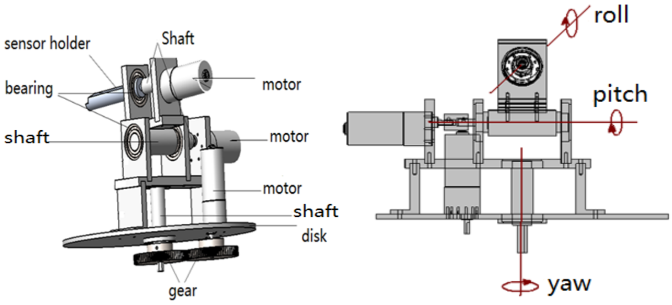

Fig. 3. The design of a 3-axis motion platform

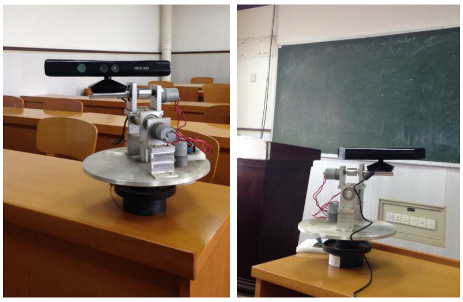

Fig. 4. A Kinect is installed upon the 3-axis motion platform

4.2 Algorithm of auto-tracking control

The axes of rotation of the 3-axis pan-tilt-roll machinery can be decomposed into three independent closed-loop control channels, and each of which consists of speed loop and position loop controlling the speed and position of the axes respectively (refer to Fig. 5) to guarantee the real-time and smooth tracking. The proportional-integral-derivative (PID) algorithm is often adopted to solve this type of control problems. The discrete-time PID controllers are expressed by the following formulas:

$$u(k) = K_P\{e(k) + \frac{T}{T_I}\sum_{i=0}^{k} e(i) + \frac{T_D}{T}[e(k) - e(k-1)]\} \tag{1}$$

where KP, TI and TD denotes the coefficient for the proportional gain, integrator gain and derivative gain, respectively; e is the difference between the set point and actual output; u is the controlled quantity applied to controlled object; T accounts for sampling period, k accounts for sampling number (k = 1,2,…,); e(i) and e(i-1) accounts for the error at the i-th and (i-1)-th sample period; u(k) accounts for the signal used to control the speed of motor by transforming into PWM signal. The three coefficients, KP, KI, and TD, normally remain unchanged once determined, and thus the traditional PID algorithm often fails to achieve optimal performance, especially when the condition changes during the operation.

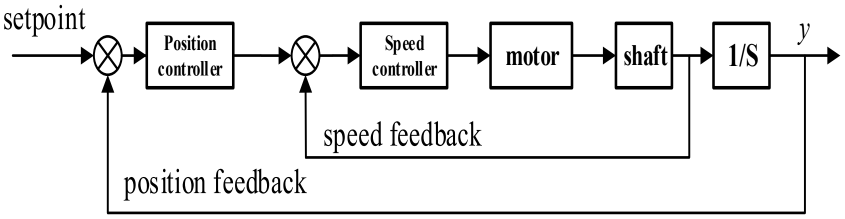

Fig. 5. Two-loop control structure with the speed loop nested inside the position loop

When Kinect detects body parts, despite the fact that present algorithms can accurately locate the joint position between shoulders, the errors are inevitable during long-time applications in actual situations. If the machinery runs based on the wrong position, it may not recover its function of tracking due to over-spinning. Furthermore, different from traditional tracking-control task, machinery should remain static when small movements are detected, to prevent frequent fibrillation from degrading the quality of video recording. Therefore, the machinery should simulate the movement manner of students' head in a natural way. Based on the above considerations, an expert PID control algorithm based on adaptive Kalman filter was implemented in our research, and the structure diagram of the control algorithm is shown in Fig. 6.

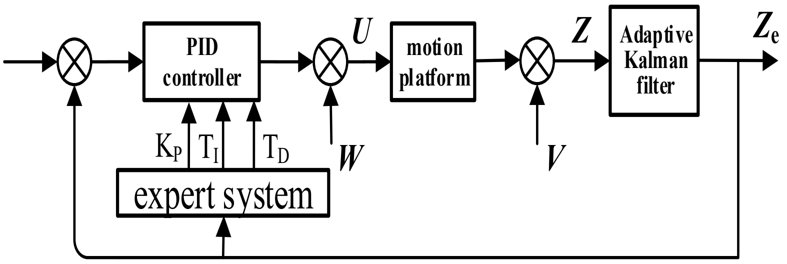

Fig. 6. Structural diagram of expert PID control algorithm based on adaptive Kalman filter

The auto-tracking system is easily affected by various interference, such as the wrong data received from the joint between shoulders by Kinect and the functionality and stability drop of PID controller. In order to reduce noise, the adaptive Kalman filter is adopted to predict basic attributes.

The discrete model for control system is as follows:

$$X_k = \Phi_{k|k-1}X_{k-1} + B_{k-1}U_{k-1} + W_{k-1}, Z_k = H_kX_k + V_k,$$

where X_k accounts for the state matrix of the system, Z_k accounts for measurement matrix; Φ accounts for state transfer matrix of the system, H accounts for measurement transfer matrix; B accounts for input matrix, U accounts for control output, W_k accounts for interference noise, V_k accounts for measurement noise.

Due to the uncertainty of the noise, this paper adopts the Sage-Husa Maximum Posterior Estimator [9] to estimate the mean value q_k and variance Q_k of W_k, and the mean value r_k and variance R_k ofV_k.

Position values of the joint between shoulders are processed by the adaptive Kalman filter, and then the calculated deviation values are sent to expert PID controller. The idea of expert PID algorithm [10] lies in all sorts of knowledge based on control theories, to modulate parameters in an intellectual way and achieve the best optimization during the process. This paper adopts two main points on positioning of the machinery: first, when the joint between shoulders goes beyond the default focus point, minimal value should be adopted; second, when the joint not only goes beyond the default focus point but also continues its movement towards the periphery, PI controller should be adopted to speed up modulating process.

Based on the specified pan-till-roll machinery and tracking algorithm, the auto-tracking on the lecturers is viable. First, depth image data acquired will be adopted by computer to run algorithm to give out the position of the joint. Second, expert PID algorithm based on Kalman filter proceeds to give out machinery control value. Finally, control data is sent to controller on the machinery by wireless serial port, and motors will be generated to power three axis to spin to assure target human body remain within Kinect’s default range. Under different lighting conditions, numerous lecture experiments are conducted, during which the auto-tracking machinery based on this paper can effectively track lecturer’s movements in real-time with relative stability. Moreover, machinery is capable of receiving orders from computer network remotely. Remote users will gain control of the machinery after given access, which at some point achieves mutual interaction in distance education.

5 Application of Intelligent Visual Sensing System in Distance Education

In this paper, intelligent visual sensing system can be applied to videotaping in distance education, as well as the platform can be extended to enable more natural human-computer interaction. Due to the limit of length, this paper only presents the framework about applications. The intelligent visual sensing system can automatically acquire multiple specific joints position such as head, arm, elbow and spine, whose motions can be used for establishing motion position recognition system. When computers recognize the specific gestures or body movements, certain command is given to switch slides, change visual angle, etc., forming a teaching system with Natural User Interface (NUI) [11, 12]. Furthermore, writing by fingertip tracking [13] and interactive remote laboratories [14] are also accessible in this platform. Kinect can also be used for face tracking (including the direction of face and facial points positioning), which can be further applied to facial expression recognition [15]. In distance education, it can be used to evaluate users' mental state[16], faces' direction for attention state, eye contact for the extent of fatigue, body movements for learning state [17]. On one hand, lecturer can receive much more feedback from users based on this particular human-computer interaction system, aiming to improve distance education itself and resolve difficulties in keeping account of users' learning state. On the other hand, with the development of VR (Virtual Reality), strong sense of realism will come soon [18], which denotes another promising aspect for distance education. The specified platform in this paper can exactly show the real world in virtual reality [19], and many other reality-enhanced teaching system can be realized. For instance, body images of users or lectures can be extracted to put into virtual reality and many other tasks like animating human movements and expressions [6]. Users' interests in learning process is greatly improved by applying this kind of natural human-computer interaction system.

In all applications mentioned above, human bodies should remain within the range of visual sensors, and stay in front to assure relatively great performance. In actual circumstances, the variance of human positions and poses and restriction of sensor's view, may not make it work as good as expected. To achieve better performance, this paper designs a visual auto-tracking machinery and its control algorithm, where more human-computer interaction schemes can be extended.

6 Conclusion

This paper adopts Kinect as the motion senor in intelligent visual recognition system and combines it with pan-till-roll machinery to achieve real-time and stable visual auto-tracking for human bodies. This particular system can not only applied in shooting distance education, but also be used as a platform where many other human-computer interaction applications can be extended as explained in this paper, which means great significance to developing and applying natural human-computer interaction technology.

7 Acknowledgement

This research is funded by the Education Research Projects No.2015A03 of the Automation Specialty Teaching Advisor Board under the Ministry of Education of P. R. China, and the Education Reform Project of Wuhan University in 2016.

8 References

[1] Li, Y., Li, L. (2013). Application of Affective Computing in Web-based Distance Education System: Functions, Current Research and Key Problems. Modern Distance Education Research, 122(2): 100-106.

[2] Li, Q., Wang, Q. (2015). Motion Sensing Technology in Education. Distance Education Magazine, (1): 48-56.

[3] Zhang, S., Qian, D. (2014). Research Status and Development Studies of Motion Sensing Technology. Journal of East China Normal University (Natural Science), (2): 40-49.

[4] Li, H., Zhao, X., Tan, Ming. (2012). Particle Filter Based Human Tracking Method in Wireless Network. Chinese Journal of Sensors and Actuators, 25(6): 807-814.

[5] Eichner, M., Marin-Jimenez, M., Zisserman, A., et al. (2012). 2D Articulated Human Pose Estimation and Retrieval in (almost) Unconstrained Still Images. International Journal of Computer Vision, 99(2): 190-214. https://doi.org/10.1007/s11263-012-0524-9

[6] Zhang, Z. (2012). Microsoft Kinect Sensor and its Effect. IEEE MultiMedia, 19(2): 4-10. https://doi.org/10.1109/MMUL.2012.24

[7] Shotton, J., Sharp, T., Kipman, A., et al. (2013). Real-time Human Pose Recognition in Parts From Single Depth Images. Communications of the ACM, 56(1): 116-124. https://doi.org/10.1145/2398356.2398381

[8] Clarke, D.W. (1984). PID Algorithms and their Computer Implementation. Transactions of the Institute of Measurement and Control, 6: 305-316. https://doi.org/10.1177/014233128400600605

[9] Sage, A.P., Husa, G.W. Algorithms for Sequential Adaptive Estimation of Prior Statistics. IEEE Symposium on 8th Decision and Control in Adaptive Processes, Nov 17-19 1969, University Park, PA, USA, pp. 61-61. https://doi.org/10.1109/SAP.1969.269927

[10] Meng, S., Liu, Y., Han, X., et al. (2014). Application of expert system PID algorithm to heat exchange station. Advanced Materials Research, 1070: 1709-1712. https://doi.org/10.4028/www.scientific.net/AMR.1070-1072.1709

[11] Villaroman, N., Rowe, D., Swan, B. Teaching natural user interaction using OpenNI and the Microsoft Kinect sensor. Proceedings of the 12th Annual ACM SIGITE Conference on Information Technology Education, New York, USA, 2011, pp. 227-232. https://doi.org/10.1145/2047594.2047654

[12] Lun, R., Zhao, W. (2015). A survey of applications and human motion recognition with Microsoft Kinect. International Journal of Pattern Recognition and Artificial Intelligence, 29(5): 1555008. https://doi.org/10.1142/S0218001415550083

[13] Feng, Z., Xu, S., Zhang, X, et al. Real-time fingertip tracking and detection using Kinect depth sensor for a new writing-in-the air system. Proceedings of the 4th International Conference on Internet Multimedia Computing and Service, Wuhan, China, 201, pp. 70-74. https://doi.org/10.1145/2382336.2382356

[14] Maiti, A. Interactive remote laboratories with gesture based interface through Microsoft Kinect. 10th International Conference on Remote Engineering and Virtual Instrumentation, Sydney, Australia, 2013, pp. 1-4. https://doi.org/10.1109/REV.2013.6502900

[15] Mao, Q., Pan, X., Zhan, Y. (2015). Using Kinect for real-time emotion recognition via facial expressions. Frontiers of Information Technology & Electronic Engineering, 16(4): 272-282. https://doi.org/10.1631/FITEE.1400209

[16] Sun, B., Chen, J., Liu, Y., et al. (2015). Application of Textual Sentiment Analysis in Individual Homework. Journal of Beijing Normal University (Natural Science), 51(3): 250-254.

[17] Zhang, H., Liu, W., Xu, W., et al. (2015). Depth Image Based Gesture Recognition for Multiple Learners. Computer Science, 42(9): 299-302.
[18] Chen, J., Huang, W., Song, A., Lu, J. Research Progress and Future Challenges of Virtual Reality Technology. Chinese Journal of Sensors and Actuators, 2002, (9): 222-227.
[19] Kyan, M., Sun, G., Li, H., et al. (2015). An approach to ballet dance training through MS Kinect and visualization in a CAVE virtual reality environment. ACM Transactions on Intelligent Systems and Technology 6(2): 1-37. https://doi.org/10.1145/2735951

9 Authors

Fei Yang, was born in April 1981. He received the PhD degree in 2010. He is currently the associate professor at the School of Engineering and Automation, Wuhan University in PR. China.

His recent research interests include robotics, intelligent sensors, image processing a nd pattern recognition.Funded by the Education Research Project No.2015A03 from th e Automation Specialty Teaching Advisory Board with the Ministry of Education of Ch ina,and the Teaching Reform Project of Wuhan University in 2016.

Rong Zhang, born in December, 1971. Received the PhD degree in 2010.Currently is a lecturer at the School of Electrical Engineering and Automation, Wuhan University in P. R. China. Recent research interests include data processing, intelligent sensors, and image processing and pattern recognition. Funded by the Teaching Reform Project of Wuhan University in 2016.

Youpeng Zhao was born on January 12, 1996. He received his Bachelor degree of Engineering in Wuhan University in June 2018. He is currently a full time graduate student in the School of Electrical and Computer Engineering, Georgia Tech. His research areas are Systems and Control, Robotics and Machine Learning.

Article submitted 2018-11-08. Resubmitted 2019-01-28. Final acceptance 2019-01-28. Final version published as submitted by the authors.