VOL. 11, NO. 11, JUNE 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ADVANCES IN THE ANALYSIS OF HUMAN GESTURE RECOGNITION USING KINECT SENSOR: A REVIEW

Vidhyapathi C. M. and Alex Noel Joseph Raj
School of Electronics Engineering, VIT University, Vellore, India
E-Mail: vidhyapathi.cm@vit.ac.in

ABSTRACT

This paper presents a comprehensive review of human gesture recognition algorithms before and after the release of low cost high resolution depth/RGB Microsoft Kinect sensor. Gesture recognition becomes an active research area in computer vision for the past one or two decades. After the release of Kinect sensor, a number of significant research advances were made particularly in pose estimation considering the depth images. Pose estimation using depth images can address major problems faced by the conventional RGB based approaches. This survey reviews the latest trends in pose estimation using depth images as well as discussing the limitations or problems faced by these approaches. The future research directions are highlighted to improve the current pose estimation algorithms. This paper expected to serve as a reference for the researchers who willing to develop new gesture recognition algorithms based on Microsoft Kinect.

Keywords: image recognition, computer vision, data fusion, Kinect sensor.

1. INTRODUCTION

There are many open problems such as identifying objects, detecting humans, activity recognition, 3D mapping in computer vision [1]. In real world setting, the reliability of object segmentation and tracking algorithms which considers only RGB images is affected by cluttered environment and illumination conditions. The availability of low cost high resolution depth and RGB camera like Kinect motivated the researchers to develop new algorithms based on RGB-D information. The complementary nature of these algorithms provides effective solution for segmentation, pose estimation, detection and tracking. The objective of the paper [2] is to summarize the overview of the approaches used in object tracking and recognition, human activity recognition, hand gesture recognition and indoor 3D mapping.

Pose estimation is the process of estimating or classifying the underlying kinematic or skeletal articulation of a person. The primary objective of any pose estimation algorithms is to propose a 3D model of skeletal joints. About 15 to 20 joints are considered for the reliable prediction or classification shown in Figure-1. Pose estimation becomes the integral step for Human activity recognition. This is a major research area because of the multi-dimensional applications right from health care, robotics to surveillance applications. More specifically gesture recognition has many applications [3]. They are sign languages, monitoring patients' emotional states, lie detection, monitoring automobile drivers and developing aids for hearing impaired.

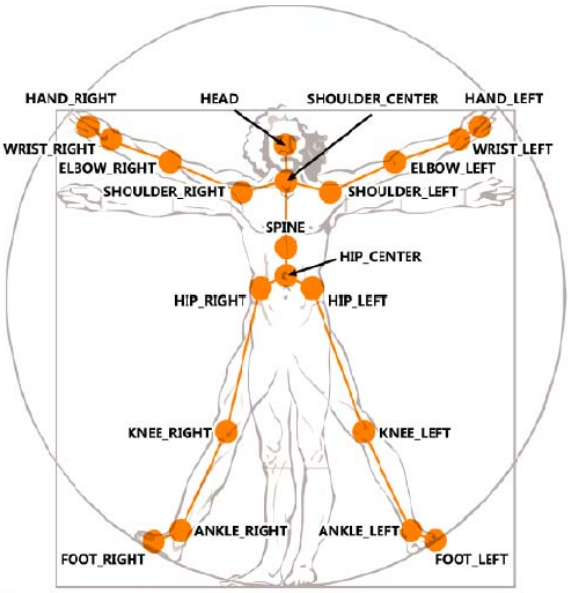

Figure-1. Kinect Skeletal joints.

Our paper compares the different types of pose estimation algorithms and its approaches in the area of human gesture recognition.

2. KINECT SENSOR HARDWARE

Kinect sensor consists of 3D Depth sensor, RGB camera and microphone array as shown in Figure-2. It can simultaneously provide the complementary nature of Depth, RGB intensity and audio output [4, 5]. Structured light 3D imaging technology is incorporated in this device [6]. The RGB camera provides 640 x 480 pixels with 8 bit per channel. It can also output higher resolution of 1280x 1024 pixels at a rate of 10 frames per second. The 3D depth sensor provides distance information from the object and the camera in the form of depth map for a range of 0.8m - 3.5m distance. The depth video is generated at a

rate of 30 frames per second with the resolution of 640x480 pixels.

Figure-2. Microsoft Kinect sensor.

Kinect provides better depth accuracy compared to ToF (Time of Flight) and Stereo camera. For short range (distance < 3.5 meters) applications Kinect outperforms TOF 3D camera and its performance is close to the ground truth data [7]. The error in the depth image increased as the distance between the object and the camera increases, ranging from a few millimetres to 4cm at the maximum range of the sensor [8]. Kinect produces minimum errors of (less than 1cm) compared to the ground truth data produced by the marker based system [9]. In a controlled environment of simple poses like standing and exercising etc., Kinect produces comparable results with the ground truth data with minimum errors. Hence Kinect can be a low cost alternative to the marker based system which is expensive. But the algorithms which used in Kinects fails for complex poses like gymnastic and acrobatic poses etc. due to occlusions and clutter with the error of 10cm [10].

3. KINECT TOOLS

OpenNI, Microsoft Kinect SDK and OpenKinect are the three mainly used tools by the research community.

3.1 OpenNI

OpenNI framework provides the library and Applications Peripheral Interface (API) to build applications using Kinect Sensor. It is a multi-platform open source framework. This frameworks works along with the middle ware called NITE. That helps the algorithm developers to build applications without much worry [11]. It is preferable to use this framework for hand gesture and hand skeletal tracking applications.

3.2 OpenKinect

OpenKinect is an open source library developed by Open community people to build applications using Kinect sensor. It also provides high level and low level Application API's to develop user applications in the area of Robotics and Human Machine Interface, Warehouse and Home automation.

3.3 Microsoft Kinect SDK

Microsoft Kinect SDK 2.0 is the latest SDK version released by Microsoft in December 2015.It is a preferred choice of developers when it comes to full body skeletal tracking without need for any pre-calibration. Itis compatible with Open NI 2.0 hence Open NI libraries are supported by Kinect SDK. Hence it is possible to develop application by taking the advantage of both packages.

4. GESTURE RECOGNITION

Gesture recognition aims to identify meaningful actions of human, considering the hands, arms, face, head and body. This is plays vital role in Human-Computer interaction. Practical implementation of gesture recognition needs tracking device or gadgets. This includes glove based gestural interface, sensor based marker tracking system, Kinect camera sensor, ToFcamera, Stereo camera or Laser sensor.

Gesture recognition can be classified as shown in Figure-3. Sign language and entertainment applications come under hand and arm gesture. Facial expressions like eyebrow rising, mouth opening, closing and moving heads are grouped under head and face gesture. Tracking the people movements in indoor and outdoor environments and analysing the full body motions are grouped with full body pose estimation.

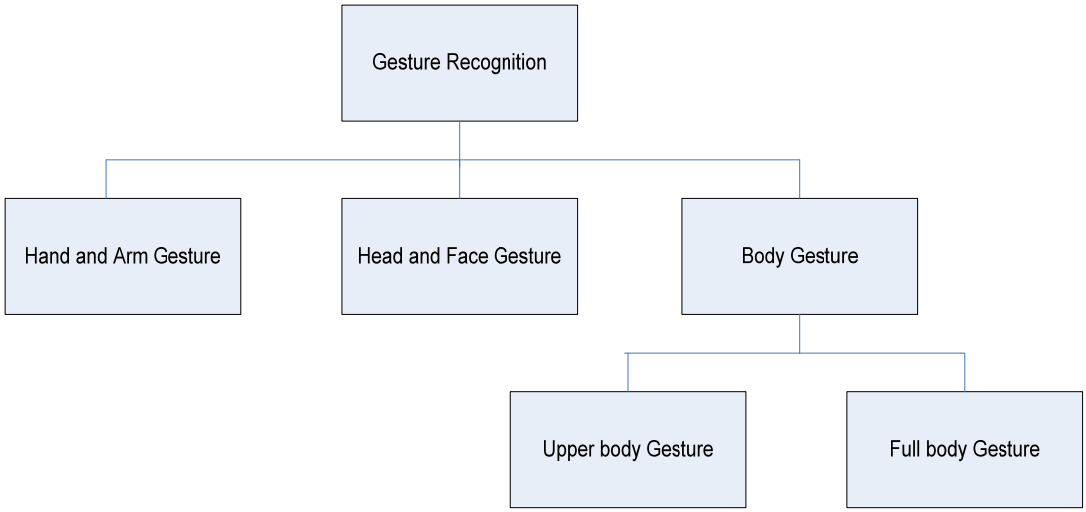

Figure-3. Type of gesture classification.

Gesture recognition approaches can be classified into different category based on the techniques adopted as shown in Figure-4.

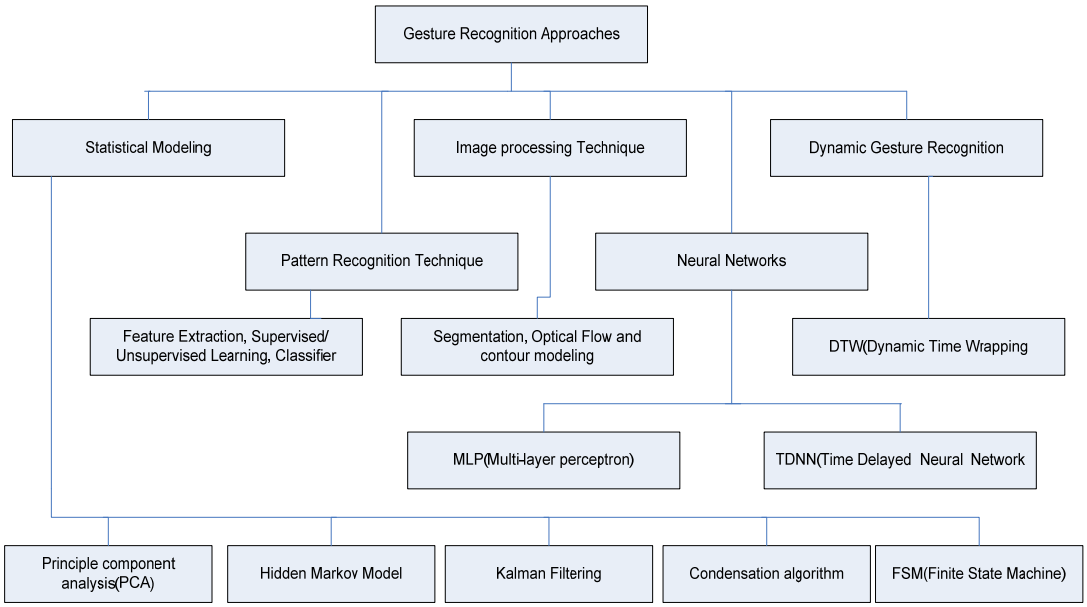

Figure-4. General classification of gesture recognition approaches..

Many of the pose estimation problems are addressed using Statistical Modelling [12-24]. The authors [25] have addressed the problem using pattern recognition technique. Image processing techniques are applied by the authors [26]. In [27] neural networks are used to address the problem. Dynamic Time Wrapping methodology can be adopted for Dynamic Gesture recognition.

4.1 Gesture recognition using RGB image

The gesture recognition algorithm can broadly be classified into three major categories. They are Model Free Approach, Indirect Model Use and Direct Model use as shown in Table-1.

Table-1. RGB image based pose estimation algorithm approaches.

Category/ [Ref]	Approaches	Limitations and Robustness
Model Free [30-34], [40], [44-45], [47-48], [50-52]	1. Probabilistic assemblies of parts(SVM classifier) 2. Example based method(HMM , NN, Nonlinear regression)	1.Pose estimation even in cluttered scenes such as sports images 2.No need of accurate priori model 3.Performs well even in single view images 4. Not view invariant, limited to fixed classes of movements
Indirect Model Use [37], [39], [41]	1. Hierarchical Rule based approach(Extended Kalman Filter) 2. Spherical Mapping Technique	1.Provides good estimate even for rapid movement 2.Requires the subjects wearing tight clothe
Direct Model Use [28, 29], [33], [34,35], [38], [42-43], [46], [49], [52-54]	1. Model based analysis by synthesis 2. Stochastic Sampling technique based on Sequential Monte Carlo 3. Learnt models of human motion 4. Deterministic gradient descent techniques 5. Extended Kalman Filter 6. Stochastic Tracking techniques/ Particle Filter 7. Deterministic grid search with gradient search 8. Stochastic meta descent 9. Stochastic sampling and search techniques. 10 Learnt Models- exemplar based approaches-3D humanoid model	1.Many of the approaches fails to predict the rapid movement 2. Particle filtering techniques can be used for whole- body pose estimation 3. These techniques fairly predicts complex poses and not much accurate compared with commercial marker based systems 4. Though these approaches perform well for specific motions, it cannot be applied for unconstrained human motion.

All the above stated gesture based algorithms are applied considering only RGB images. When it comes to real world setting, RGB image based pose estimation schemes are not reliable. This is mainly because the environment is cluttered or the illumination may change over time, both of which frequently occur in real world situation. Also many of these algorithms fail due to rapid movement and they produce less accuracy for complex poses. The price and poor quality of depth cameras restricted the researchers to work only with RGB images.

4.2 Gesture recognition using Kinect sensor

Microsoft has released its first low cost high resolution depth and RGB camera in 2010. The complementary nature of depth and RGB information has been used effectively to address many conventional problems such as illumination, cluttering, occlusions etc., faced by the pose estimation algorithms. After the release of Kinect sensor many researchers have started developing new algorithms and/or new applications based on Depth and RGB information.

Table-2. Different scheme of Kinect pose estimation.

Type of information used from Kinect	Number of Kinect Sensors used	Difficulty level and challenging in generating 2D model	Difficulty level and challenging in generating 3D model	Estimated or classified skeletal accuracy/Robustness	Computational complexity
Single View Depth Image	Single	Medium	More	Less	Less
Multiple View Depth Image	Single or Multi	Less	Medium	Medium	Medium
RGB and Depth Image	Single or Multi	Less	Medium	High	High

There are three types of research trends are identified from the literature and are presented in the Table-2. Developing the 3D skeletal models from a single view depth image has been identified as a difficult problem to address.

Table-3. RGB and Depth image based pose estimation algorithm approaches.

Approach	Single/Multiple depth or RGB image	Output	Real time implementation	Limitations/Robustness	Classifier	Data set used	Full body/upper body/Hand tracking
Identify individual body parts [55]	Single Depth Image	3D joint model	Yes/5ms per frame	Invariant to pose, body shape and clothing Computational efficiency	Randomised Decision Forest	Synthetic depth image dataset	Full body
Regression directly from depth image[56]	Single Depth Image	Multiple 3D joints	Yes/200 frames per second	Ability to localize occluded and visible body parts.	Regression and Classification function/Regression forest	MSRC-5000 pose estimation test set	Full body
One shot pose estimation [57]	Depth and Multiview Silhouette images	3D joint Model	Yes	Invariant to body size and shape	Random Forest	Synthetic depth image dataset	Full body
Matching with pre-captured motion exemplars [58]	Single depth image	3D pose estimation	No	View independent, handles body size variation, Average accuracy of 38mm	No	Synthetic depth image dataset	Full body
Exemplar based approach [59]	Single Depth image	3D skeletal points	Yes	Focus more on Pose correction	Random Forest Regressors	Own data set generated by Kinect sensor	Full body
Supervised Pose classification [60]	RGB and Depth Image	Recognised poses	Yes	Focus on correct classification	Support Vector machine	Own data set	Full body
Based on Geodesic distances between different points on the body[61]	RGB and Depth image	Skeleton body model	Yes	Ability to track poses with self-occlusions , higher accuracy	No	No	Full body
Part based hand gesture recognition [62]	Depth and RGB image	Recognise hand gesture	Yes	93.2% accuracy with the efficiency of 0.0750, Robustness to distortions and hand variations	No	Own dataset containing 1000 cases	Hand Gesture Recognition
Pose recognition in parts [63]	Single Depth Image	3D positions of body joints	Yes	0.731 to .677 mAP(Average precsion), Runs in less than 5ms	Randomised Decision Forest	Synthetic shilhouette images and own data set	Full body

Many authors [55, 56, 58, and 59] have attempted to address this problem and have succeeded in their approaches with few limitations as shown in Table-3. The algorithms are reported to be computationally efficient but they fail for rapid movement and complex poses.

Thi-Lan *et al.* [61] has attempted the problem considering both the complementary nature of Depth and RGB information which inturn improves the overall accuracy and robustness but with increased computational complexity.

5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The introduction of low cost high resolution Depth/RGB Kinect sensor has motivated the research in the field of Human Computer Interaction (HCI) in developing reliable pose estimation algorithms. The pose estimation algorithms developed using depth information can effectively to overcome the drawbacks of conventional RGB based algorithms. We believe that the introduction of Kinect surely has brought us close to the goal. But still there are number of problems in developing reliable and computationally efficient gesture recognition algorithm as listed below.

5.1 Information Fusion architecture

There are lot of scope in developing pose estimation algorithms considering Depth and (RGB) visual information. Information fusion architecture can be proposed to effectively combine complementary nature of Depth and RGB information. This would help us to

achieve better accuracy and robustness compared to the schemes which consider only RGB or Depth Images.

5.2 Implementation on FPGA (Field Programmable Gate array) or GPU (Graphical Processing Unit)

In many of the algorithms, the RGB and Depth images are processed separately as two streams of data. There are many common module used in the processing pipeline of RGB and Depth information. Identifying common modules and developing new algorithms can avoid duplication in the process. Instead of feeding all the information to all the algorithmic modules, selectively feeding the needed information can also improves the computational efficiency. As many of the steps in the algorithm can be processed in parallel, implementing the algorithm on an FPGA (Field Programmable Gate Array) or in GPU (Graphical Processing Unit) would make the algorithms to run in real time.

REFERENCES

[1] Jungong Han., Ling Shao., Dong Xu. and Jamie Shotton. 2013. Enhanced Computer Vision with Microsoft Kinect Sensor: A Review. IEEE Transactions on Cybernetics. 43(5).

[2] Ling shao., Jungong Han., Dong Xu., and Jamie Shotton. 2013. Computer Vision for RGB-D Sensors: Kinect and its applications. IEEE Transactions on Cybernetics. 43(5).

[3] Susmitha Mitra. and Tinku Acharya. 2007. Gesture Recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications And Reviews. 37(3).

[4] Tashev I. 2011. Recent advances in human-machine interfaces for gaming and entertainment. Int. J. Inform. Technol. Security. 3(3): 69-76.

[5] Kumatani K., Arakawa T., Yamamoto. K., *et al*. 2012. Microphone array processing for distant speech recognition: Towards real-world deployment. in Proc. APSIPA Annu. Summit Conf. pp. 1-10.

[6] Geng J. 2001. Structured-light 3-D surface imaging: A tutorial. Adv. Optics Photonics. 3(2): 128-160.

[7] Smisek J., Jancosek M. and Pajdla T. 2011. 3-D with Kinect. In: Proc. IEEE ICCV Workshops. pp. 1154-1160.

[8] Khoshelham K. and Elberink S. 2012. Accuracy and resolution of Kinect depth data for indoor mapping applications. Sensors. 12(2): 1437-1454.

[9] Dutta T. 2012. Evaluation of the Kinect sensor for 3-D kinematic measurementin the workplace. Appl. Ergonom. 43(4): 645-649.

[10] Obdrzalek S., Kurillo G., Ofli F., et al. 2012. Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population. in Proc. IEEE EMBC. pp. 1188-1193.

[11] Jae-Han Park., Yong-Deuk Shin, Ji-Hun Bae, et al. 2012. Spatial uncertainty model for visual features using Kinect sensor, Sensors.

[12] Rabiner L, R. 1989. A tutorial on hidden Markov models and selected applicationsin speech recognition. Proc. IEEE. 77(2): 257-285.

[13] Yamato J., Ohya J. and Ishii K. 1992. Recognizing human action in times equential images using hidden Markov model. In: Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn., Champaign, IL. pp. 379-385.

[14] Samaria F. and Young S. 1994. HMM-based architecture for face identification. Image Vis. Comput. 12: 537-543.

[15] Welch G. and Bishop G. 2000. An introduction to the Kalman filter. Dept. Comput. Sci., Univ. North Carolina, Chapel Hill, Tech. Rep. TR95041.

[16] Arulapalam S., Maskell S., Gordon N., et al. 2001. A tutorial on particlefilters for on-line nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. 50(2): 174-188.

[17] Kwok C., Fox D. and Meila M. 2004. Real-time particle filters. Proc. IEEE. 92(3): 469-484.

[18] Isard M. and Blake A. 1996. Contour tracking by stochastic propagation ofconditional density. in Proc. Eur. Conf. Comput. Vis., Cambridge, U.K. pp. 343-356.

[19] Michael isard, Andrew Blake. 1998. Condensation-Conditional density propagation for visual tracking. Int. J. Comput. Vis. 1: 5-28.

[20] Doucet A., de Freitas N. and Gordon N. 2001. Eds., Sequential Monte Carloin Practice. NewYork: Springer-Verlag.

[21] Davis J. and Shah M. 1994. Visual gesture recognition. Vis., Image Signal Process. 141: 101-106.

[22] Bobick A, F. and Wilson A, D. 1997. A state-based approach to the representation and recognition of gesture. IEEE Trans. Pattern Anal. Mach. Intell. 19(12): 1235-1337.

[23] Yeasin M. and Chaudhuri S. 2000. Visual understanding of dynamic handgestures. Pattern Recogn. 33: 1805-1817.

[24] Hong P., Turk M. and Huang T, S. 2000. Gesturemodeling and recognitionusing finite state machines. In: Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recogn, Grenoble, France. pp. 410-415.

[25] Gonzalez R, C. and Woods, R, E. 1992. Digital Image Processing. Reading,MA: Addison-Wesley.

[26] Kass M., Witkin A. and Terzopoulos D. 1988. SNAKE: Active contour models. Int. J. Comput. Vis. pp. 321-331.

[27] Haykin S. 1994. Neural Networks: A Comprehensive Foundation. New York: Macmillan.

[28] Agarwal A., Triggs B. 2006. Recovering 3D human pose from monocularimages. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28(1): 44-58.

[29] Aggarwal J, K., Cai Q. 1999. Human motion analysis: a review, ComputerVision and Image Understanding. 73(3): 428-440.

[30] Aggarwal J, K., Cai Q., Liao W., et al. 1994. Articulated and elasticnon-rigid motion: a review. in Workshop on Motion of Non-Rigid and Articulated Objects, Austin, Texas.

[31] Aggarwal J, K., Cai Q., Liao W., et al. 1998. Nonrigid motionanalysis: articulated and elastic motion. Computer Vision and Image Understanding. 70(2): 142-156.

[32] Aggarwal J, K., Park S. 2004. Human motion: modeling and recognition of actions and interactions. in: Second International Symposium on 3D Data Processing, Visualization and Transmission, Thessaloniki, Greece.

[33] Ahmad M., Lee S. 2006. Human action recognition using Multiview image sequence features. in: International Conference on Automatic Face and Gesture Recognition, Southampton, UK.

[34] Allen B., Curless B., Popovic Z. 2002. Articulated body deformation from range scan data. in: ACM SIGGRAPH. pp. 612-619.

[35] Ambrosio J., Abrantes, J., Lopes G. 2001. Spatial reconstruction of human motion by means of a single camera and a biomechanical model. Journal of Human Movement Science. 20: 829-851.

[36] Ambrosio J., Lopes G., Costa J., et al. 2001. Spatial reconstruction of the human motion based on images of a single camera. Journal of Biomechanics. 34: 1217-1221.

[37] Andersen P, F., Corlin R. 2005. Tracking of Interacting People and Their Body Parts for Outdoor Surveillance. Master's thesis, Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark.

[38] Antonini G., Martinez S, V., Bierlaire M., et al. 2006. Behavioral priors for detection and tracking of pedestrians in video sequences. International Journal of Computer Vision. 69(2): 159-180.

[39] Arikan O., Forsyth D, A. 2002. Synthesizing constrained motions from examples. in: ACM SIGGRAPH. pp. 483-490.

[40] Atsushi N., Hirokazu K., Shinsaku H., et al. 2002. Tracking multiple people using distributed vision systems. in: International Conference on Robotics and Automation, Washington DC, USA.

[41] Azoz Y., Devi L., Yeasin M. 2003. Tracking the human arm using constraint fusion and multiple-cue localization. Machine Vision and Applications. 13(5-6): 286-302.

[42] Babu R,V., Ramakrishnan K, R. 2003. Compressed domain human motion recognition using motion history information. In: International Conference on Acoustics, Speech and Signal Processing, Hong Kong, China.

[43] Balan A, O., Black M, J. 2006. An adaptive appearance model approach for model-based articulated object tracking. in: Computer Vision and Pattern Recognition, New York City, New York, USA.

[44] Balan A, O., Sigal L., Black M, J. 2005. A quantitative evaluation of video based 3D person tracking, in: Workshop on Visual Surveillance and

Performance Evaluation of Tracking and Surveillance, Beijing, China.

[45] Barron C., Kakadiaris I, A. 2000. Estimating anthropometry and pose from a single image, in: Computer Vision and Pattern Recognition. Hilton Head Island, South Carolina.

[46] Barron C., Kakadiaris I, A. 2001. Estimating anthropometry and pose from a single uncalibrated image. Computer Vision and Image Understanding. 81(3): 269-284.

[47] Barron C., Kakadiaris I, A. 2003. The improvement of anthropometry and pose estimation from a single uncalibrated image. Machine Vision and Applications 14(4): 229-236.

[48] Beleznai C., Fruhstuck B., Bischof H. 2005. Tracking multiple humans using fast mean shift mode seeking. in Workshop on Performance Evaluation of Tracking and Surveillance, Breckenridge, Colorado.

[49] Belongie S., Malik J., Puzicha J. 2001. Matching shapes. in International Conference on Computer Vision, Vancouver, Canada.

[50] Ben-Arie J., Wang Z., Pandit P., *et al*. 2002. Human activity recognition using multidimensional indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence. 24(8): 1091-1104.

[51] Ben Abdelkader C., Cutler R., Davis L. 2002. Motion-based recognition of people in EigenGait space. in: International Conference on Automatic Face and Gesture Recognition, Washington DC, USA.

[52] Berclaz J., Fleuret F., Fua P 2006. Robust people tracking with global trajectory optimization. in: Computer Vision and Pattern Recognition, New York City, New York, USA.

[53] Billard A., Epars Y., Calinon S., *et al*. 2004. Discovering optimal imitation strategies. Robotics and Autonomous Systems. 47: 69-77.

[54] Bissacco A., Soatto S. 2006. Classifying human dynamics without contact forces. In: Computer Vision and Pattern Recognition, New York City, New York, USA.

[55] Shotton J., Fitzgibbon A., Cook *et al*. 2011. Real-time human pose recognition in parts from a single depth image. In: Proc. IEEE Conf. Comput. Vision Pattern Recognit. pp. 1297–1304.

[56] Girshick R., Shotton J., Kohli P., et al. 2011. Efficient regression of general-activity human poses from depth images. in Proc. ICCV. pp. 415-422.

[57] Taylor J., Shotton J., Sharp T., *et al.* 2012. Thevitruvian manifold: Inferring dense correspondences for one-shot human pose estimation', in Proc. CVPR. pp. 103-110.

[58] Ye M., Wang X., Yang R., et al. 2011. Accurate 3-D pose estimation from A single depth image. in Proc. ICCV. pp. 731-738.

[59] Shen W., Deng K., Bai X., *et al*. 2012. Exemplarbasedhuman action pose correction and tagging. in Proc. IEEE Conf. Comput. Vision Pattern Recognit. pp. 1784-1791.

[60] Thi-Lan Le., Minh-Quoc Nguyen., Thi-Thanh-Mai. Human Posture Recognition using human skeleton provided by Kinect. Nguyen international Research Institute Michhust-Cnrs/Umi-2954-Grenoble Inphanoi University of Science and Technology Vietnam. 978-1-4673-2088/13, IEEE

[61] Loren Arthur Schwarz, Artashes Mkhitaryan., Diana Mateus, *et al*. 2012. Human skeleton tracking from depth data using geodesic distances and optical flow. Image and Vision Computing. 30: 217-226.

[62] Zhou Ren., Junsong Yuan., Jingjing Meng., *et al*. 2013. Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. IEEE Transactions on Multimedia. 15(5).

[63] Jamie Shotton., Alex Kipman., Andrew Fitzgibbon., *et al*. 2013. Real-Time Human pose recognition in Parts from Single Depth Images. Communications of the ACM. 56(1).