SCIENCE CHINA
Information Sciences

• MOOP •

September 2016, Vol. 59 093101:1–093101:5
doi: 10.1007/s11432-016-5604-y

A concise tutorial on human motion tracking and recognition with Microsoft Kinect

Wenbing ZHAO

Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, Ohio 44115*, USA*

Received April 30, 2016; accepted May 25, 2016; published online August 23, 2016

Abstract This paper provides a concise tutorial on the Microsoft Kinect technology and the state of art research on human motion tracking and recognition with Microsoft Kinect. A pre-requisite for human motion recognition is feature extraction. There are two types of feature extraction methods: skeleton joint based, and depth/color image based. Given a set of feature vectors, a motion could be recognized using machine learning, direct comparison, or rule-based methods. We also outline future research directions on the Kinect technology.

Keywords Microsoft Kinect, depth camera, human motion tracking and recognition, machine learning

Citation Zhao W B. A concise tutorial on human motion tracking and recognition with Microsoft Kinect. Sci China Inf Sci, 2016, 59(9): 093101, doi: 10.1007/s11432-016-5604-y

1 Introduction

Microsoft Kinect is a revolutionary depth camera first released in late 2010. Kinect is a milestone in the areas of computer vision in that it drastically increases the accessibility of depth sensors. Previously, depth sensors often cost on the order of several thousand US dollars. Furthermore, Microsoft provides excellent programming support for the Kinect sensor, and offers a license for the commercial use of its software development kit (SDK) free of charge. Compared with traditional cameras, the special meaning of Kinect is that it offers highly accurate depth sensing. This is coupled with the excellent and free SDK, which enables the development of Kinect-based applications for developers who do not have deep computer vision background to develop various applications to solve real-world problems.

In less than five years, we have seen the use of Microsoft Kinect in a wide spectrum of applications, ranging from medicine and healthcare, to education and performing arts, to robotics, to sign language recognition, and many others. The foundation for these applications rests on the human motion tracking and recognition using the information provided by Kinect.

The accuracy of human motion recognition heavily depends on the extraction of the most distinctive features of the motion. The Kinect technology significantly reduces the difficulty of this crucial step by providing a skeleton stream that contains full-body joint positions and orientations. Software developers could focus on the recognition task using machine learning, direct comparison, or rule-based methods based on features derived from the joint data.

Email: wenbing@ieee.org

In this paper, we first provide a brief overview of the Kinect technology and a concise tutorial on the state of the art research on human motion tracking and recognition with Kinect. For a far more comprehensive review of this subject, readers are referred to [1]. We also outline some future research on the Kinect technology.

2 The Microsoft Kinect technology

The Kinect sensor is equipped with a color camera and a depth sensor that is capable of measuring the depth of each pixel in its view. One of the most significant advancement made by the Kinect SDK is that an application could register to receive skeleton frames, in addition to color and depth frames. The skeleton tracking technology used in Kinect was based on the research carried out by Microsoft Research UK [2].

There are two main versions of Microsoft Kinect. The initial Kinect was released for Xbox 360, Microsoft's game controller, in November 2010. In February 2012, a slightly revised version, referred to as Kinect for Windows, was released. Kinect for Windows has the added functionality of near-mode upper-extremity human motion tracking. Both the original Kinect and Kinect for Windows use the same depth sensing technology, hence, they are regarded as the same version and are referred to as Kinect v1. In summer 2014, Kinect for XboxOne (Microsoft's latest game console) was released. The new Kinect is referred to as Kinect v2 because it uses a completely different depth sensing technology. Microsoft also made drastic changes on its SDK for Kinect v2, and it is incompatible with that for Kinect v1.

In Kinect v1, depth sensing is based on structured light, which can be considered as a form of triangulation. The technology was developed by PrimeSense, and it makes it possible to use a single infrared (IR) emitter and a depth sensor to compute the depth of each pixel. The structured light pattern (emitted by the IR emitter) enables the depth sensor to derive the line from the IR emitter to each pixel. Because the distance between the IR emitter and the depth sensor is known, the depth of the pixel can be computed easily.

In Kinect v2, the depth sensing technology was abandoned due to its low fidelity and a time-of-flight technology was employed. The depth of each pixel is computed based on the phase shift of the emitted modulated light and the corresponding reflected light.

3 Human motion tracking and recognition

The ultimate goal of human motion tracking is to understand the semantics of the motion, i.e., what the motion means. Human motion can be roughly categorized into two classes: gestures and activities. A gesture consists of primarily the movement of one or two hands and it typically conveys some specific meaning to another person for inter-person interaction, or to a machine (i.e., computer/game console/robot) for human-machine interaction. An activity typically involves the movement of many segments of the body, such as walking, running, brushing teeth. Physical exercises and sports activities are also examples of human activities. For an activity, the type to which the activity belongs can be considered as its semantics. Sometimes, additional semantics such as the fine-grained characteristics of the activity, are of interest. For example, for both physical exercises and sports activities, there are specific stipulations on how the activities should be performed.

To recognize a human gesture or activity, the first task is to determine the most distinctive set of features of the motion. This step is referred to as feature extraction. The next step is to understand the semantics of the motion based on the set of features extracted.

3.1 Feature extraction

With the availability of a full skeleton joint positions and orientations for each user tracked, the task of feature extract becomes much simpler. One only need to determine which joints are the most important for each activity. For a gesture, the two hand positions are almost always used as features. Nevertheless,

there are substantial research works with Kinect that extract features directly from the depth frames, or both depth and color frames. These methods can be useful when Kinect skeleton frames are not available, or not accurate enough. For example, Kinect runtime could perform skeleton segmentation only when a user is within a specific range.

The methods used for feature extraction from the depth/color frames are mostly adapted from traditional feature extraction methods from video frames. The additional depth information has been proven to help extract more distinctive features to recognize activities that are difficult based on video frames. The features are first extracted from every frame (depth or color). If both color and depth frames are used, the features are fused together. Then a clustering algorithm, such as kmeans, is used to derive a set of most distinctive feature vectors. These feature vectors form a visual vocabulary, often referred to as a codebook, which would be used for classification of unknown human motions [3].

3.2 Motion recognition

Machine learning methods are the most popular ways of performing human motion recognition. For simpler activities and gestures, algorithmic (i.e., rule-based) methods are often used. For offline recognition, direct comparison based methods could also be used.

3.2.1 *Algorithmic-based recognition*

In algorithmic-based recognition, a set of rules for key poses and the sequencing of the poses are defined for each gesture or activity. Once the rules are available, the implementation of recognition algorithms can be easily implemented with very high recognition accuracy and speed. To accommodate Kinect measurement errors and intrinsic tolerance of the motion, error bounds are typically necessary and they would have to be tuned for each user. This recognition method is popular in healthcare and gaming applications because human motions in these contexts are often well defined and relatively simple [4, 5].

3.2.2 *Direct comparison*

In this approach, a database of human motions is first built using training data. Then, given an unknown motion, various methods are employed to compare the new motion with each in the database, and the unknown motion is classified as the one with the minimum distance between the two. The most popular method is dynamic time warping, which could compute the similarity between two temporal sequences despite timing and speed differences [6].

3.2.3 *Machine learning*

In this approach, sophisticated statistical models, such support vector machine (SVM), decision trees/forests, the hidden Markov model (HMM), and artificial neural networks (ANNs), are used to capture the unique characteristics of a gesture or an activity. These models often consist of a significant number of parameters. Some of the parameters must be tuned manually. Other parameters will be determined automatically using training data.

SVM is the most popular method for classification based on codebooks [3]. SVMs are supervised learning models that aim to achieve maximum-margin separation using relatively little training data. Training data is used to determine a plane (for linear classification) or hyperplane (for nonlinear classification) that separates the data that belong to different classes as further away as possible. This plane or hyperplane then can be used to classify unknown data into different classes.

Decision trees/forests fit the task of gesture recognition well if a gesture can be modeled as a sequence of key poses. The path from a leaf node to the root note can then be used to classify unknown gestures with trained decision trees/forests where each internal node represents a key pose [7].

HMM is another widely used model for motion recognition because it captures the temporal characteristics of a gesture/activity explicitly. Similar to decision trees/forests, HMM also fits well when a gesture/activity can be modeled as a sequence of poses [8]. Different from decision trees/forests, however,

the poses that are key to a gesture/activity do not have to be known explicitly, i.e., they can be learned using training data. For more sophisticated activities, hierarchical HMMs have been used to model sub-activities as well as the composite activities [9]. Although HMM often requires large amount of training data and recognition could be slow, we have seen research work that uses HMM for one-shot learning and realtime recognition [10].

4 Future research directions

The Kinect technology can be made even more useful by integrating it with other modalities of motion sensing. This would enable a seamless integration of indoor and outdoor tracking of the activities of daily lives for both the general population as well as demographics that require frequent monitoring by healthcare professionals, such as post-injuries and post-surgery patients, people with disabilities, and the elderlies [11]. The integration of the two modalities also facilitates selective user tracking using Kinect, i.e., the Kinect can selectively track those who have explicitly consented to be tracked, which preserves the privacy of those who are also present in the view of the Kinect sensor [12, 13].

Obviously, there is still a large room to improve on the automatic recognition of human activities using Kinect [14]. Selecting feature vectors for the recognition is critical to the recognition accuracy, and yet it is usually manually done. Deep neural networks could potentially help in the selection of the most critical features for highly reliable recognition.

5 Conclusion

In this article, we presented a concise tutorial on the Kinect technology and the state of the art human motion tracking and recognition in Kinect applications. We categorized feature extraction methods into two types: (1) skeleton joint based; and (2) depth/color image based, which would employ sophisticated computer vision algorithms. The latter is more powerful but much more expensive than the former. There are three major approaches to human motion recognition: (1) machine learning based; (2) direct comparison based; and (3) rule based. Furthermore, we outlined some future research directions of the Kinect technology.

Acknowledgements This work was supported in part by a Faculty Research Development award, and a Graduate Faculty Travel award, both from Cleveland State University.

Supporting information The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Lun R, Zhao W B. A survey of applications and human motion recognition with microsoft kinect. Int J Pattern Recogn Artif Intell, 2015, 29: 1555008
2 Shotton J, Sharp T, Kipman A, et al. Real-time human pose recognition in parts from single depth images. Commun ACM, 2013, 56: 116–124
3 Xia L, Chen C C, Aggarwal J. View invariant human action recognition using histograms of 3d joints. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, 2012. 20–27
4 Zhao W B, Lun R, Espy D D, et al. Realtime motion assessment for rehabilitation exercises: integration of kinematic modeling with fuzzy inference. J Artif Intell Soft Comput Res, 2014, 4: 267–285
5 Zhao W B, Lun W, Espy D D, et al. Rule based realtime motion assessment for rehabilitation exercises. In: Proceedings of the IEEE Symposium on Computational Intelligence in Healthcare and e-Health, Orlando, 2014. 133–140
6 Masood S, Qureshi M P, Shah M B, et al. Dynamic time wrapping based gesture recognition. In: Proceedings of the International Conference on Robotics and Emerging Allied Technologies in Engineering, Islamabad, 2014. 205–210

7 Miranda L, Vieira T, Martinez D, et al. Real-time gesture recognition from depth data through key poses learning and decision forests. In: Proceedings of the 25th SIBGRAPI Conference on Graphics, Patterns and Images, Ouro Preto, 2012. 268–275
8 Xu D, Chen Y L, Lin C, et al. Real-time dynamic gesture recognition system based on depth perception for robot navigation. In: Proceedings of the IEEE International Conference on Robotics and Biomimetics, Guangzhou, 2012. 689–694
9 Sung J, Ponce C, Selman B, et al. Unstructured human activity detection from RGBD images. In: Proceedings of the IEEE International Conference on Robotics and Automation, Saint Paul, 2012. 842–849
10 Zamborlin B, Bevilacqua F, Gillies M, et al. Fluid gesture interaction design: applications of continuous recognition for the design of modern gestural interfaces. ACM Trans Interact Intell Syst, 2014, 3: 22
11 Foti D, Kanazawa L. Activities of daily living. In: Pendleton H M, Schultz-Krohn W, eds. Pedretti's Occupational Therapy: Practice Skills for Physical Dysfunction. St. Louis: Mosby, 2008. 146–194
12 Zhao W B. On automatic assessment of rehabilitation exercises with realtime feedback. In: Proceedings of the IEEE International Electro-Information Technology Conference, Grand Forks, 2016. 376–381
13 Zhao W B, Espy D D, Reinthal M A, et al. Privacy-aware human motion tracking with realtime haptic feedback. In: Proceedings of the IEEE International Conference on Mobile Services, New York, 2015. 446–453
14 Zhao W B, Lun R, Gordon C, et al. A privacy-aware kinect-based system for healthcare professionals. In: Proceedings of the IEEE International Electro-Information Technology Conference, Grand Forks, 2016. 205–210