Virtual Boutique: customizable gesture recognition with Kinect Sensor

Di Wu, Fan Zhu, Ling Shao
The University of Sheffield
Sheffield, UK
{elp10dw,fan.zhu,ling.shao}@sheffield.ac.uk

Abstract

Hand gesture based Human-Computer-Interaction (HCI) is one of the most natural and intuitive ways to communicate between people and machines and the one-shot learning scenario is one of the real life situations in terms of gesture recognition problems. In this demo, we present a hand gesture recognition system using the Kinect senor, which addresses the problem of one-shot learning gesture recognition with a user-defined training and testing system. Such system can behave as a remote controller while the user can allocate a specific function using a preferenced gesture by performing it only once. To adopt the gesture recognition framework, the system first automatically segments an action sequence into atomic tokens, and the Extended-Motion-History-Image (Extended-MHI) *is used for motion feature representation. We evaluate the performance of our system quantitatively in Chalearn Gesture Challenge [1], and apply it to a virtual clothing shopping system.*

1. Introduction

Automated sign language recognition from video has been studied for at least about twenty years. Hands convey a significant amount of information including configurations, positions and instantaneous velocities. However, the recognition of continuous, natural signing remains challenging due to their high deformability, large number of degrees of freedom and high level of self-occlusion, which gives rise to an enormous variation of appearance and a high level of ambiguity.

For most traditional machine learning problems, typical algorithms are based on large amounts of training data. However, for some real-life machine learning tasks, obtaining such large amounts of training data is not always possible. Thus, one-shot learning algorithms that aim to learn information from one, or only a handful, training samples are brought on the table for discussion. Previous presented methods [2] tackle with the one-shot learning problem by transferring prior knowledge into the new learning task. Such methods mimic the way how human are believed to recognize new objects and work quite well for object recognition. However, it is sometimes too expensive to collect sufficient previous knowledge for simple recognition tasks.

In this demo, we propose a simple functional gesture recognition system with user-defined gestures according to the users preference. Such user-defined gesture settings make our system a privilege for the use of people who have certain physical disabilities or when they are not at their convenience in performing certain gestures. The learning of our system bases only on the new incoming samples, without leveraging previous data obtained from other sources.

2. System overview and performance

The framework of our system is shown in Figure 1. Due to the existence of some imperfection/noise of various sources in current depth sensor [4], a spatial filtering and a morphological preprocessing step are adopted for depth image noise reduction. In order to make our system effectively respond to the user's consecutive gestures, the existence of a rest position is ruled to segregate the consecutive gestures, where the rest position is also user-defined. For the above temporal segmentation task, we adopt a simple but effective appearance based approach [7] to retrieve similar frames and a non-maximum suppression approach is used to trim the borderline of the action tokens. Taking advantage of the unique property of the kinect depth sensor [3], human gesture silhouettes are segmented from the backgrounds, upon which the Extended-Motion-History-Image *(Extended-MHI)* [6] technique is applied as a global representation. Finally, a non-parametric *maximum-correlation-coefficient* classifier is adopted to circumvent the issue of overfitting.

We quantify our recognition rate by computing the Levenshtein distance ($\mathfrak{LD}$) between the list of predicted labels R and the corresponding list of true labels T on the dataset of [1] and achieve the less than 0.26 in $\mathfrak{LD}$ on the final batch of the dataset, which is very promising for real life applications.

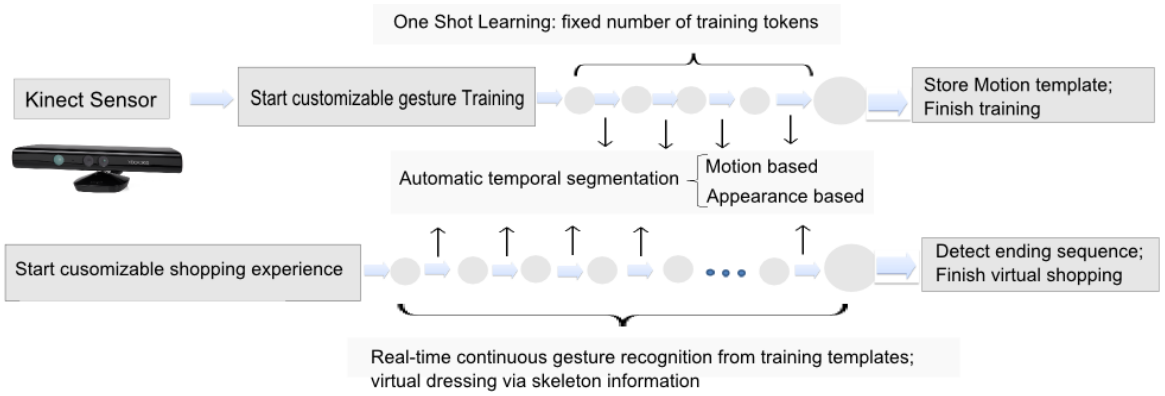

Figure 1. Demonstration of system flowchart.

3. Applications

The goal of this demo is to qualitatively showcase the robustness of our hand gesture recognition system on the top of our algorithms applied in [7]. For real-life gesture recognition machines, a tailored gesture vocabulary is necessary for a certain task. With the help of one-shot learning, a new set of vocabulary can be easily generated before executing a new task. In our system, we adopt a "customizable" framework and apply it to the "Virtual Boutique" application, which can be used to enhance the shopping experience.

As in Figure 2, at the training phase, customized gesture samples are captured to generate the new gesture vocabulary with only one training sample. Different gesture samples are automatically segmented using both motion and appearance information. At the testing phase, the same temporal segmentation framework is similarly adopted as in the training phase. Then, virtual boutique listens to shopper's command, *e.g.*, "Yes, I like it", "No, I hate it", "Next item","Put it on", "Take it off", "Show 3D", "End shopping", *etc*. Using the body joins algorithm from [5], we attire the online shopper accordingly. The end of shopping session is determined by the detection of "End shopping" gesture token. Our system performs less than 0.2 in $\mathfrak{LD}$ given sufficient motion input.

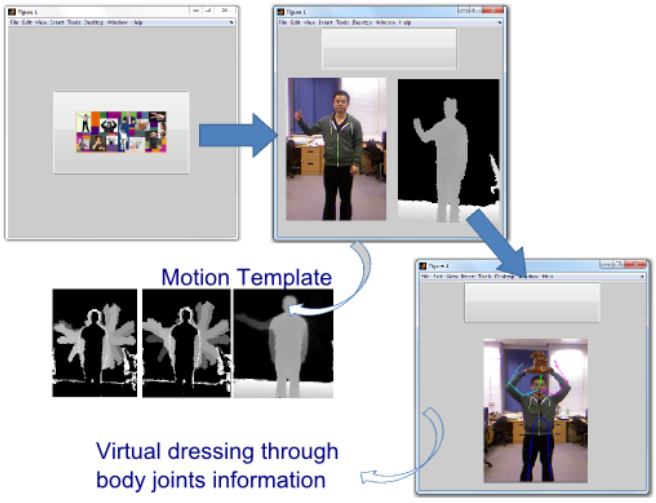

Figure 2. Demonstration of Virtual Boutique.

4. Conclusion

In this demo, we present a one-shot learning gesture recognition system via the "Virtual Bouteque". where customizable gestures can be recognized through the input of a Kinect sensor. Both the depth and color information obtained from the Kinect sensor are used for action representation, which ensures the robustness of our system to cluttered environments. Besides, the proposed system recognizes new gestures with only one training example per gesture class. Such a system provides a robust solution in real-life HCI applications and many other hand gesture based HCIs.

References

[1] Chalearn gesture dataset. *CGD2011, ChaLearn, California*, 2011. 1

[2] R. F. L. L. Fei-Fei and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. 1

[3] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun. Real time motion capture using a single time-of-flight camera. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010. 1

[4] J. K. S. Park, S. Yu and S. Lee. 3d hand tracking using kalman filter in depth space. *EURASIP Journal on Advances in Signal Processing*, 2012. 1

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 2

[6] D. Wu and L. Shao. Silhouette analysis based action recognition via exploiting human poses. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2012. 1

[7] D. Wu, F. Zhu, and L. Shao. One shot learning gesture recognition from rgbd images. In *In CVPR2012 workshop on gesture recognition*, 2012. 1, 2