Virtual Try-on using Kinect and HD camera

Stevie Giovanni[⋆], Yeun Chul Choi[§], Jay Huang[§], Eng Tat Khoo[§], KangKang Yin[⋆]

[⋆]National University of Singapore {stevie,kkyin}@comp.nus.edu.sg
[§]EonReality {yeunchul.choi,jay.huang,engtat}@eonreality.com

Abstract. We present a virtual try-on system - EON Interactive Mirror - that employs one Kinect sensor and one High-Definition (HD) Camera. We first overview the major technical components for the complete virtual try-on system. We then elaborate on several key challenges such as calibration between the Kinect and HD cameras, and shoulder height estimation for individual subjects. Quality of these steps is the key to achieving seamless try-on experience for users. We also present performance comparison of our system implemented on top of two skeletal tracking SDKs: OpenNI and Kinect for Windows SDK (KWSDK). Lastly, we discuss our experience in deploying the system in retail stores and some potential future improvements.

Keywords: virtual try-on, Kinect, HD camera, OpenNI, Kinect for Windows SDK

1 Introduction

The fashion industry greatly relies on traditional retail outlets because most users only feel comfortable purchasing outfits after physically trying them out. The consequences of this fact include that Internet shopping is hard for clothing; and fitting rooms in brick-and-mortar stores are always packed during peak hours. This motivates us in developing a virtual try-on system that enables shoppers to digitally try out clothes and accessaries. Our system, named EON Interactive Mirror(http://www.eonreality.com/), utilizes one Kinect sensor and one High-Definition (HD) camera. The Interactive Mirror enables the shopper to virtually try-on clothes, dresses, handbags and accessories using gesture-based interaction. Customers experience an intuitive and fun way to mix-and-match collections without having to queue for fitting rooms or spend time changing items. Customers can also snap pictures of their current selections and share them on Social Media to get instant feedback from friends, which can potentially shorten the decision time for making the purchase.

Our system has been deployed since April 2012 in one of Singapore's largest shopping centers with approximately three million visitors passing through every month. With EON Interactive Mirror, walk-by customers can be convinced to walk into the store. Within the store, it has created unique customer experiences of virtually trying on the latest fashion 'on-the-go' in a fun and engaging way, and made the store stand out from the highly competitive market.

In order to achieve a believable virtual try-on experience for the end user, several challenges have to addressed. First, the Kinect sensor can only provide low-resolution

VGA quality video recording, yet high quality video is essential for attractive visual appearance on large screens. We thus opt to use an HD camera to replace the role of Kinect's built-in RGB camera. This necessitates a calibration process between the HD camera and the Kinect depth camera in order to map the 3D clothes seamlessly to the HD video recording of the customers. Second, digital clothes need to be resized to fit to a user's body. Yet the Kinect depth data is noisy, the skeletal motion tracked by third-party SDKs is not accurate, and sometimes the lower part of the body is not even in the camera's field of view. We thus need a reliable and robust procedure to estimate the shoulder height of the users for the clothes-body fitting process.

We will first discuss related previous work in Section 2. We then give an overview of our virtual try-on system in Section 3, followed by details of the above mentioned key components in ensuring a seamless user experience. Skeletal motion tracking is implemented on two SDKs; and their performance comparison is documented in Section 4. Lastly, we present our observation and experience in deploying our system for retail store customers, and discuss the limitations of the current system for potential future improvements.

2 Related Work

Markerless human motion tracking is a long-standing problem in computer vision. With the recent advances in depth cameras and sensors, especially the Kinect sensor [2], research on human skeletal pose tracking has made great improvements [5, 11, 12, 14, 15]. Our system builds on top of these techniques by utilizing publicly available SDKs that incorporate some of these state-of-the-art algorithms.

Kinect has also enabled various interactive applications that are creative and fun, such as ShoeSense [6], MirageTable [7], HoloDesk [8], FaceShift [13], TeleHuman [10], and KinectFusion [9]. Most relevant to our Interactive Mirror is the ever-growing virtual fitting room systems available on the market, such as Fitnect [1] and TriMirror [4]. However, we have not been able to find any technical details of these systems. From their demo videos alone, the major difference between our system and TriMirror, for example, is that we do not simulate clothes in our system. We simply render the deformed clothes on top of the user's video stream, and this requires a high-quality calibration between the Kinect and the HD camera.

3 System Overview

Our virtual try-on system consists of a vertical TV screen, a Microsoft Kinect sensor, an HD camera, and a desktop computer. Fig. 1 shows the front view of the Interactive Mirror together with the Kinect and HD camera. The Kinect sensor is an input device marketed by Microsoft, and intended as a gaming interface for Xbox 360 consoles and PCs. It consists of a depth camera, an RGB camera, and microphone arrays. Both the depth and the RGB camera have a horizontal viewing range of 57.5 degrees, and a vertical viewing range of 43.5 degrees. Kinect can also tilt up and down within -27 to +27 degrees. The range of the depth camera is [0.8$\sim$4]m in the normal mode and [0.4$\sim$3]m in the near mode. The HD camera supports a full resolution of 2080×1552, from which

Fig. 1: *The front view of the Interactive Mirror with Kinect and HD camera placed on top.*

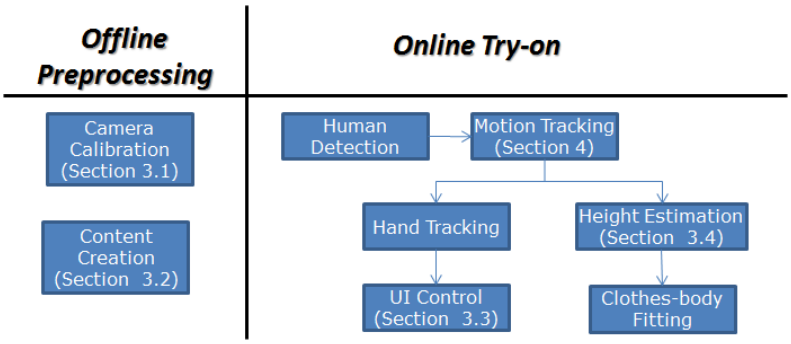

Fig. 2: *Major software components of the virtual try-on system.*

we crop out the standard HD resolution 1920 × 1080. It supports a frame rate of 60Hz with a USB 3.0 interface of up to 5 Gbit/s transfer rate. The Interactive Mirror is a 65" TV screen mounted in portrait mode with HD resolution 1920 × 1080. We recommend a space of [1.5∼2.5]m × [2.0∼2.5]m × [2.0∼2.5]m (width×legth×height) in front of the mirror as the virtual fitting room.

Fig. 2 illustrates the major software components of the virtual try-on system. During the offline preprocessing stage, we need to calibrate the Kinect and HD cameras, and create 3D clothes and accessories. These two components will be discussed in more details in Sections 3.1 and 3.2 respectively. During the online virtual try-on, we first detect the nearest person among the people in the area of interest. This person will then become the subject of interest to be tracked by the motion tracking component implemented on two publicly available Kinect SDKs, as will be discussed in Section 4. The user interacts with the Interactive Mirror with her right hand to control the User Interface (UI) and select clothing items. The UI layout will be discussed in more details in Section 3.3.

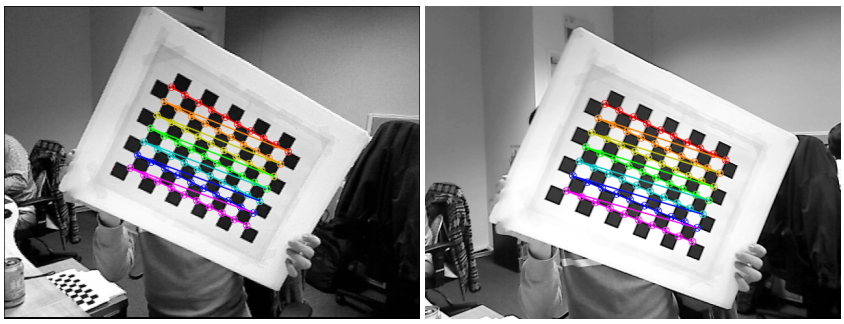

Fig. 3: *The camera calibration process. The checkerboard images seen by the Kinect RGB camera (left) and the HD camera (right) at the same instant of time.*

For good fitting of the clothes onto the body, we need to estimate the height of the user to resize the digital clothes appropriately. We discuss two ways of height estimation in Section 3.4. The ratio between the height of the real user and that of the default digital model will then be used to scale the clothes uniformly in three dimensions. Finally, the resized digital clothes are skinned to the skeleton, rendered with proper camera settings, and merged with the video stream of the user.

3.1 Camera calibration

Vision-based augmented reality systems need to trace the transformation relationship between the camera and the tracking target in order to augment the target with virtual objects. In our virtual try-on system, precise calibration between the Kinect sensor and the HD camera is crucial in order to register and overlay virtual garments seamlessly onto the 2D HD video stream of the shoppers. Furthermore, we prefer a quick and semi-automatic calibration process because the layout between Kinect and HD camera with respect to the floor plan may be different for different stores, or even for the same store at different times. To this end, we use the *CameraCalibrate* and *StereoCalibrate* modules in OpenCV [3] for camera calibration. More specifically, we recommend to collect a minimum of 30 pairs of checkerboard images seen at the same instant of time from Kinect and HD camera, and calculate each pair's correspondences, as shown in Fig. 3.

In addition, the Kinect sensor is usually not perfectly perpendicular to the ground plane, and its tilting angle is needed to estimate the height of users later in Section 3.4. We simply specify the floor area from the Kinect depth data manually, and the normal vector of the floor plane in Kinect's view can be calculated. The tilting angle of Kinect is then the angle between this calculated floor normal and the gravity normal.

Furthermore, to seamlessly overlay the virtual garments on top of the HD video, we also need to estimate the tilting angle of the HD camera, and a correct FoV (Field of View) that matches the TV screen's aspect ratio. Subsequently precise perspective transformations can be applied by our rendering engine to properly render the deformed digital clothes for accurate merging with the HD video.

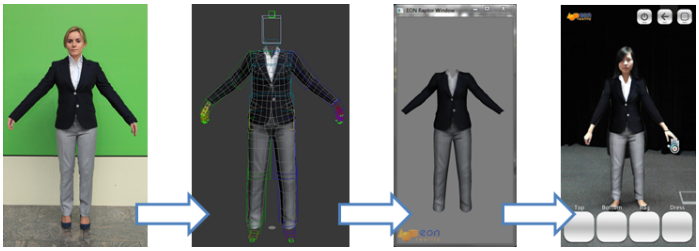

Fig. 4: *Major steps for content creation. Catalogue images are first manually modeled and textured offline in 3DS Max. We then augment the digital clothes with relevant size and skinning information. At runtime, 3D clothes are properly resized according to a user's height, skinned to the tracked skeleton, and then rendered with proper camera settings. Finally, the rendered clothes are merged with the HD recording of the user in realtime.*

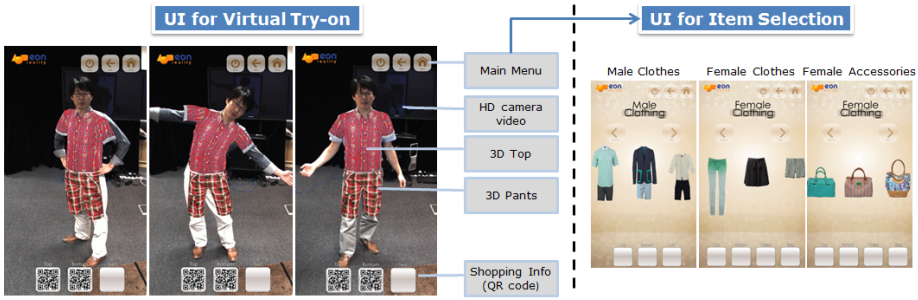

Fig. 5: *Left: the UI for virtual try-on. Right: the UI for clothing item selection.*

To summarize, the output of the camera calibration procedure include:

- extrinsic camera parameters (translation and rotation) of the HD camera with respect to the Kinect depth camera.
- the tilting angles of the Kinect sensor and the HD camera with respect to the horizontal ground plane.
- FoV of the HD camera.

3.2 Content creation

Our virtual 3D clothes are based on actual catalogue images, so that new fashion lines can be added to the system quickly. Fig. 4 shows the major steps of converting catalogue images to 3D digital clothes. In the preprocessing stage, our artists manually created one standard digital male mannequin and one female mannequin. Then they modeled the catalogue images into 3D clothes that fit the proportions of the default mannequins. Corresponding textures were also extracted and applied to the digital clothes. Then we augment the digital clothes with relevant size and skinning information. At runtime, 3D

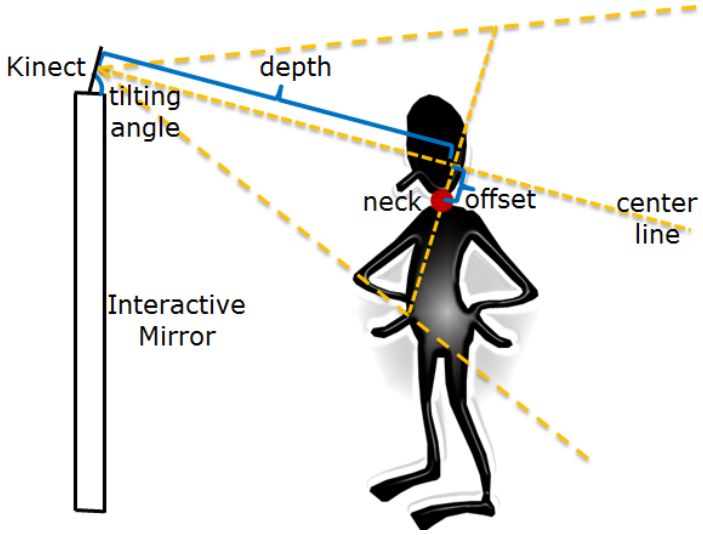

Fig. 6: *Shoulder height estimation when the user's feet are not in the field of view of Kinect. The tilting angle of the Kinect sensor, the depth of the neck joint, and the offset of the neck joint with respect to the center point of the depth image can jointly determine the physical height of the neck joint in the world space.*

clothes are properly resized according to a user's height, skinned to the tracked skeleton, and then rendered with proper camera settings. Lastly, the rendered clothes are merged with the HD recording of the user in realtime.

Our content development team modeled 115 clothing items in total, including male clothes, female clothes, and accessories. On average it took about two man days to create and test one item for its inclusion into the virtual try-on system.

3.3 User interface

Fig. 5 depicts the user interface of the Interactive Mirror. Because our clothes are 3D models rather than 2D images, users are able to turn their body within a reasonable range in front of the Interactive Mirror and still have the digital clothes properly fit to their body, just like what they can see in front of a real mirror. The user selects menu items and outfit items using hand gestures. Different tops, bottoms, and accessories can be mixed and matched on the fly.

3.4 Height estimation

Digital clothes need to be rescaled according to users' body size, for good fitting and try-on experiences. We propose two methods to estimate a user's shoulder height. The first one simply uses the neck to feet height difference, when both the neck and the feet joints are detected by Kinect skeletal tracking SDKs. As illustrated in Fig. 6, however,

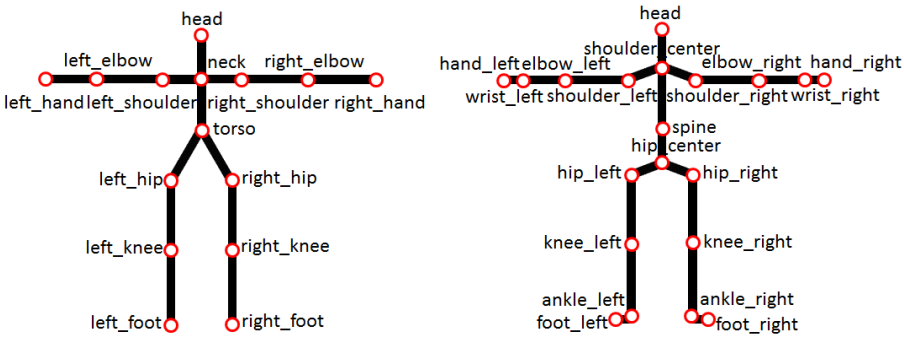

Fig. 7: *Human skeletons defined by OpenNI (left) and KWSDK (right).*

sometimes the feet are not located within the field of view of Kinect. In such scenarios, we can still estimate the neck height from the tilting angle of the Kinect sensor, the depth of the neck joint in the Kinect depth image, and the offset of the neck joint with respect to the center point of the depth image. After the shoulder/neck height is estimated, we then uniformly resize the digital clothes in three dimensions for a better fit to the user's body.

4 Skeletal Motion Tracking: OpenNI vs. KWSDK

One key component of a virtual try-on system is to track the motion of the user. We built our motion tracking component on the successful Kinect sensor and publicly available SDKs developed for Kinect. More specifically, we have experimented with two SDKs that provide skeletal motion tracking capability for Kinect. The first one is the OpenNI 1.1.0.41. OpenNI is a set of open source SDKs released by an organization of the same name. It aims to standardize applications that access natural interaction devices. The other SDK we use is Kinect for Windows SDK (KWSDK) 1.5, released by Microsoft to support developers who wish to work with Kinect. Here we begin with an overview of both SDKs, and then we compare their performance related to skeletal motion tracking.

Both OpenNI and KWSDK can query the Kinect sensor for RGB images and depth images up to 30 frames per second. Additionally, both can track a user's skeleton that includes information of positions and orientations of each joint. Their major difference lies in the structure of the returned skeletons, shown in Fig. 7. Note that in OpenNI the neck joint always lies on the line that connects the left and right shoulders, while the KWSDK shoulder_center joint does not necessarily lie on the shoulder line.

OpenNI requires a skeleton calibration step before it can track user's poses; while KWSDK can work in a walk in/walk out situation. On the other hand, KWSDK is more prone to false positives, such as detecting chairs as users. In addition, KWSDK cannot correctly identify the right vs. left limbs of the user when she faces backwards away from Kinect. This problem is depicted in Fig. 8. The red lines represent the limbs on the left side of the tracked skeleton.

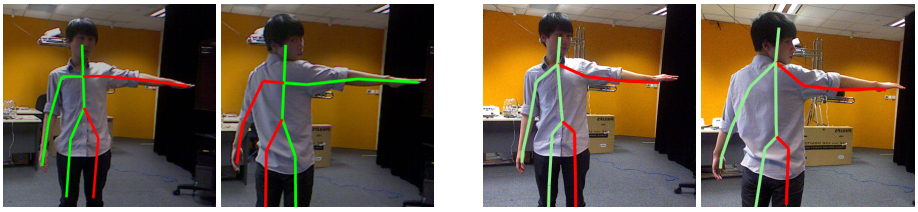

Fig. 8: *Left: OpenNI correctly identifies the left limbs (colored red) regardless the facing direction of the user. Right: yet KWSDK confuses the left and right limbs when the user faces backwards.*

measured shoulder height (cm)	neck-to-feet OpenNI (cm)	shoulder_center-to-feet KWSDK (cm)	neck height OpenNI (cm)	shoulder_center height KWSDK (cm)
153.4	134.6	153.2	156.8	162.7
151.0	129.5	149.7	153.5	161.8
151.0	116.5	136.0	149.0	158.8
144.2	114.4	141.3	148.2	151.2
143.5	121.3	139.0	147.5	146.0
143.5	117.1	138.9	147.3	148.2
137.6	105.4	131.6	143.7	142.9
135.5	105.0	129.3	142.0	135.6
134.0	106.1	129.0	142.1	137.8

Table 1: *Comparison of shoulder height estimation between OpenNI and KWSDK. Column 1: manual measurements; Column 2&3: height estimation using neck to feet distance; Column 4&5: height estimation using the method of Fig. 6 when feet positions are not available.*

In addition to full-body skeletal tracking, OpenNI provides functionalities such as hand tracking, gesture recognition, background foreground separation etc. KWSDK supports additional capabilities such as seated skeletal tracking, face tracking, speech recognition, background separation etc. Our system currently does not utilize these features and components.

4.1 Performance Comparison

We first compare the performance of OpenNI and KWSDK in terms of their joint tracking stability. To this end, we recorded 30 frames (1s) of skeleton data from three subjects holding the standard T-pose standing from various distances (1.5m, 2m, and 2.5m) to the Kinect sensor. The Kinect was placed 185cm above the ground and tilted downward 20 degrees. Subject 1 and 2 were males wearing a polo or T-shirt, jeans, and casual sneakers, of height 190.5cm and 173cm respectively. Subject 3 was a 163cm female wearing a blouse, jeans, and flat flip-flops. We then calculated the standard deviation of each joint position for all the visible joints. Fig. 9 shows the results, which suggest that the joint tracking stability of OpenNI and KWSDK are roughly comparable. Note that

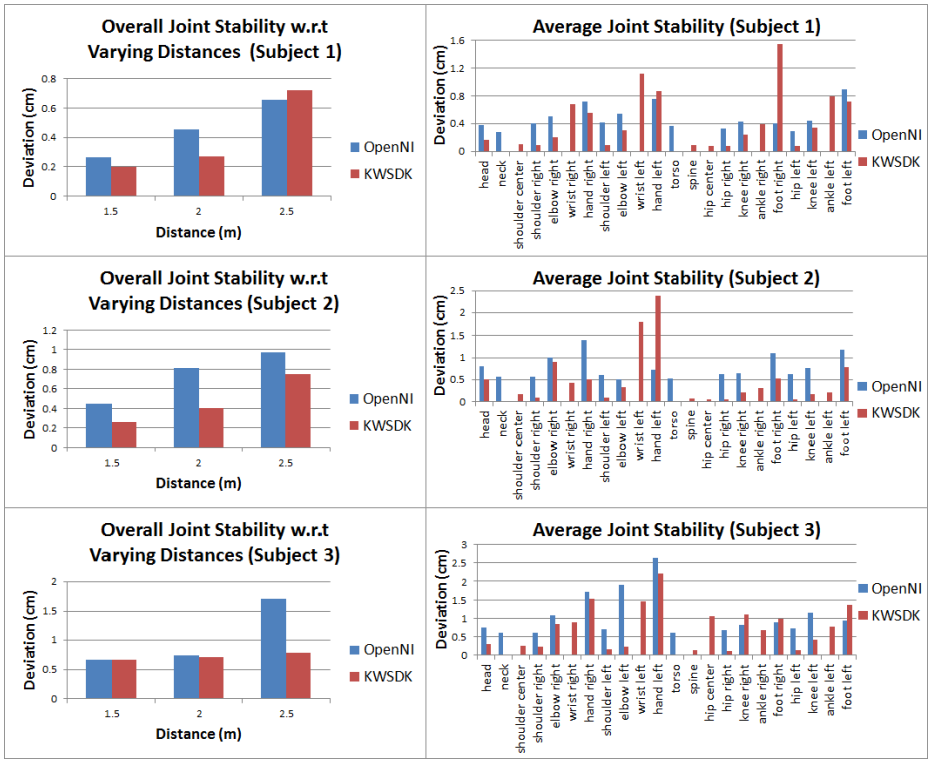

Fig. 9: *Comparison of joint tracking stability between OpenNI and KWSDK. Left: average standard deviation of joint positions in centimeters for all joints across all frames. Right: average standard deviation for each individual joint across all frames.*

we recorded the T-pose trials with KWSDK while doing online tracking. We then fed the recorded depth data to OpenNI to do offline tracking. Thus the same T-pose trials were used for both SDKs to eliminate the difference caused by users' motion variations. Ideally, we should also capture the same trials using a high-end motion capture system such as Vicon, so that the joint tracking stability of the two SDKs from Kinect data can be compared with ground truth data. Due to space and time constraints, however, we did not perform such comparison. From the average individual joint stability charts in the right column of Fig. 9, we can also see that end-effectors such as hands and feet are more unstable compared to inner body joints in both SDKs.

We also compare how OpenNI and KWSDK integrate with our height estimation methods described in Section 3.4. With Kinect placed 185cm above the ground and titled down 20 degrees, we captured nine subjects wearing T-shirts and jeans and holding the T-Pose for one second two meters away from the mirror. At this distance, the subject's full-body skeleton could be seen. We first simply calculated the average distance from the neck joint (in OpenNI) or shoulder_center joint (in KWSDK) to the mid-point of the feet joints as shoulder height estimation. The results are shown in the second and

third columns of Table 1. Second, we used the method depicted in Fig. 6 for height estimation without using the feet positions. The results are shown in the fourth and fifth columns of Table 1. The first column of Table 1 lists our manual measurement of the vertical distance between the floor to the mid-point of the clavicles. This is the shoulder height that our clothes-body fitting algorithm expects to overlay the virtual clothes. We can see from Table 1 that feet-to-neck heights tend to underestimate the shoulder heights, mainly because there is usually a distance between the feet and the ground that is not compensated for by the first height estimation method. For the second approach that does not use feet positions, such underestimation is eliminated. On the other hand, KWSDK tends to overestimate the height now, mainly because its shoulder_center joint usually locates above the shoulder line, as shown in Fig. 7 right.

5 Discussion

EON Interactive Mirror offers several advantages over traditional retailing. It attracts more customers through providing a new and exciting retail concept, and creates interest in the brand and store by viral marketing campaigns through customers sharing their experiences in Social Media such as Facebook. Furthermore, it reduces the need for floor space and fitting rooms, thereby reducing rental costs and shortening the time for trying on different combinations and making purchase decisions. We encourage interested readers to search our demo videos with keywords EON Interactive Mirror at http://www.youtube.com.

We have closely engaged participating retailers during the content creation process in an iterative review process to ensure the high quality of interactive 3D clothes from catalog images. Thus the retailers and shopping mall operators were confident and excited to feature their latest fashion lineups with EON Interactive Mirror. The try-on system was strategically placed in the high traffic flow area of the shopping mall, and successfully attracted many customers to try on the virtual clothes and bags. The retailers appreciated the value of the system as a crowd puller, and to allow other passers-by to see the interaction when somebody is trying clothes with the Interactive Mirror. We have also observed that interactions with the system were often social, where either couples or group of shoppers came together to interact with the mirror. They took turns to try the system, and gave encouragement when their friend or family was trying. Notably the system also attracted families with young children to participate. In this case, the parents would assist the children in selecting the clothes or bags. Due to limitations of Kinect SDKs, the system would not be able to detect or has intermittent tracking for children shorter than one meter. However, this limitation did not stop the young children from wanting to play with the Mirror.

Currently there are several limitations of our system. First, the manual content creation process for 3D clothes modeling is labor intensive. Automatic or semi-automatic content creation, or closer collaboration and integration with the fashion design industry will be needed to accelerate the pace of generating digital clothing for virtual try-on applications. Additionally, our current clothes fitting algorithm scales the outfit uniformly. This is problematic when the user is far way from the standard portion. For instance, a heavily over-weighted person will not be covered entirely by the virtual clothes because

of her excessive width. Extracting relevant geometry information from the Kinect depth data is a potential way to address this problem.

In the future, we wish to augment the basic try-on system with an additional recommendation engine based on data analytics, so that the system could offer customers shopping suggestions 'on the fly' regarding suitable sizes, styles, and combinations to increase sales of additional clothes or promote matching accessories. The system could also be used to gather personalized shopping preferences, and provide better information for market research on what create just an interest to try versus a decision to buy. We would also like to explore the possibility of adapting our system for Internet shopping, for customers who have a Kinect at home. In this scenario, we will simply use the RGB camera in Kinect rather than an additional HD camera.

Acknowledgements: This work was partially supported by Singapore Ministry of Education Academic Research Fund Tier 2 (MOE2011-T2-2-152).

References

1. Fitnect. `http://www.fitnect.hu/`
2. Kinect for Windows. `http://www.microsoft.com/en-us/kinectforwindows/`
3. Opencv. `http://opencv.org/`
4. Trimirror. `http://www.youtube.com/user/triMirrorTV`
5. Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: IEEE 13th International Conference on Computer Vision (ICCV). pp. 1092–1099 (2011)
6. Bailly, G., Müller, J., Rohs, M., Wigdor, D., Kratz, S.: Shoesense: a new perspective on gestural interaction and wearable applications. In: CHI'12. pp. 1239–1248 (2012)
7. Benko, H., Jota, R., Wilson, A.D.: Miragetable: Freehand interaction on a projected augmented reality tabletop. In: CHI'12 (2012)
8. Hilliges, O., Kim, D., Izadi, S., Weiss, M., Wilson, A.D.: Holodesk: Direct 3d interactions with a situated see-through display. In: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems. CHI '12 (2012)
9. Izadi, S., Newcombe, R.A., Kim, D., Hilliges, O., Molyneaux, D., Hodges, S., Kohli, P., Shotton, J., Davison, A.J., Fitzgibbon, A.: Kinectfusion: real-time dynamic 3d surface reconstruction and interaction. In: SIGGRAPH 2011 Talks. p. Article 23 (2011)
10. Kim, K., Bolton, J., Girouard, A., Cooperstock, J., Vertegaal, R.: Telehuman: effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod. In: CHI'12. pp. 2531–2540 (2012)
11. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N.: Estimating human 3d pose from time-of-flight images based on geodesic distances and optical flow. In: IEEE Conference on Automatic Face and Gesture Recognition (FG). pp. 700–706 (2011)
12. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR. pp. 1297–1304 (2011)
13. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Trans. Graph. 30(4), Article 77 (2011)
14. Ye, M., Wang, X., Yang, R., Ren, L., Pollefeys, M.: Accurate 3d pose estimation from a single depth image. In: ICCV. pp. 731–738 (2011)
15. Zhu, Y., Dariush, B., Fujimura, K.: In: CVPRW'08, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops