Human Behavior Analysis from Depth Maps

Sergio Escalera[1,2]

[1] Dept. Matemàtica Aplicada i Anàlisi,
Universitat de Barcelona, Gran Via de les Corts Catalanes 585,
08007, Barcelona, Spain
[2] Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Spain
sergio@maia.ub.es

Abstract. Pose Recovery (PR) and Human Behavior Analysis (HBA) have been a main focus of interest from the beginnings of Computer Vision and Machine Learning. PR and HBA were originally addressed by the analysis of still images and image sequences. More recent strategies consisted of Motion Capture technology (MOCAP), based on the synchronization of multiple cameras in controlled environments; and the analysis of depth maps from Time-of-Flight (ToF) technology, based on range image recording from distance sensor measurements. Recently, with the appearance of the multi-modal RGBD information provided by the low cost Kinect™ sensor (from RGB and Depth, respectively), classical methods for PR and HBA have been redefined, and new strategies have been proposed. In this paper, the recent contributions and future trends of multi-modal RGBD data analysis for PR and HBA are reviewed and discussed.

Keywords: Pose Recovery, Human Behavior Analysis, Depth Maps, Kinect™.

1 Introduction

Pose Recovery (PR) uses to be a first step of most Human Behavior Analysis (HBA) systems. However, detecting humans and recovering their pose in images or videos is a challenging problem due to the high variety of possible configurations of the scenario (such as changes in the point of view, illumination conditions, or background complexity) and the human body (because of its articulated nature). In the past few years, some research on PR has focused on the use of Time-of-Flight range cameras (ToF) [1–4]. Nowadays, several works related to this topic have been published because of the emergence of inexpensive structured light technology, reliable and robust to capture the depth information along with their corresponding synchronized RGB image. This technology has been developed by the PrimeSense [5] company and released to the market by Microsoft® XBox® under the name of Kinect™.

F.J. Perales, R.B. Fisher, and T.B. Moeslund (Eds.): AMDO 2012, LNCS 7378, pp. 282–292, 2012.

With the recent wide use of the depth maps introduced by the Microsoft® Kinect™ device, a new source of information has emerged. With the use of depth maps, 3D information of the scene from a particular point of view is easily computed, and thus, working with consecutive frames, we obtain RGBDT information, from Red, Green, Blue, Depth, and Time data, respectively. This motivates the use of multi-modal data fusion strategies to benefit from the new data representation in PR and HBA applications. While these tasks could be achieved by inter-frame feature tracking and matching against predefined gesture models, there are scenarios where a robust segmentation of human limbs are needed, e.g. observing upper limb anomalies or distinguishing between finger configurations while performing a gesture. In that respect, depth information appears quite handy by reducing ambiguities due to illumination, colour, and texture diversity. Many researchers have obtained their first results in the field of human motion capture using this technology. In particular, Shotton et al. [6] presented one of the greatest advances in the extraction of the human body pose from depth images, an approach that also is the core of the Kinect™ human recognition framework. Moreover, new devices offering multimodal RGBD + audio information are appearing [7], improving the 320×240 and 640×480 resolution of Kinect™ Depth and RGB images, consolidating the field of research, and opening the possibilities for a new broad range of applications.

Currently, there exists a steady stream of updates and tools that provide robustness and applicability to the device. In December 2010, OpenNI [8] and PrimeSense [5] released their own Kinect™ open source drivers and motion tracking middleware for PCs runningWindows (7, Vista, and XP), Ubuntu, and MacOSX. Then, the middleware FAAST (Flexible Action and Articulated Skeleton Toolkit) was developed at the University of Southern California (USC) Institute for Creative Technologies to facilitate the integration of full-body control within virtual reality applications and video games when using OpenNI-compliant depth sensors and drivers [9, 10]. In June 2011, Microsoft® released a non-commercial Kinect™ Software Development Kit (SDK) forWindows that includes Windows 7-compatible PC drivers for the Kinect™ device [11]. Microsoft® SDK allows developers to build Kinect™ enabled applications in Microsoft® Visual Studio 2010 using C++, C# or Visual Basic. Microsoft® has released a commercial version of the Kinect™ for Windows SDK with support for more advanced device functionalities. There is also a third set of Kinect™ drivers for Windows, Mac and Linux PCs by the OpenKinect (libFree- Nect) open source project [12], adapted by libraries commonly used on Computer Vision as OpenCV. Code Laboratories CL NUI Platform offers a signed driver and SDK for multiple Kinect™ devices on Windows XP, Vista, and 7 [13]. As a consequence of the new data representation obtained from Microsoft® Kinect™, new libraries to process depth maps have emerged, such as the Point Cloud Library (PCL) [14].

Some examples of applications that have benefited from RGBD representation are: reconstruction of dense surfaces and 3D object detection [15], improved descriptors and learning for object recognition [16, 17], augmented reality [18], SLAM [19], or PR-HBA, just to mention a few. In this paper, recent literature on PR and HBA using depth maps is reviewed. Once PR is robustly performed using RGBD representation, standard techniques for HBA can be consequently improved. HBA is extremely challenging because of the huge number of possible configurations of the human body that defines human motion. Common approaches to model sequential data for gesture recognition are based on Hidden Markov Model (HMM) [20], which consist of learning the transition probabilities among different human state configurations, and, more recently, there has been an emergent interest in Conditional Random Field (CRF) [21] for the learning of sequences. However, all these methods assume that we know the number of states for every motion. Other approaches make use of templates or global trajectories of motion, being highly dependent of the environment where the system is built. In order to avoid all these situations, Dynamic Time Warping framework (DTW) [22] allows to align two temporal sequences taking into account that sequences may vary in time based on the subject that performs the gesture. The alignment cost can be then used as a gesture appearance indicator. In comparison to classical 2D approaches, the authors of [23] show that an improved PR description based on 3D skeletal model from RGBD data allows to compute feature variability, and include this measure in DTW for improved HBA recognition.

The rest of the paper is organized as follows: Section 2 reviews the most recent achievements and proposals in PR based on depth maps to improve the accuracy of standard HBA systems and Section 3 concludes the paper, discussing future trends in this field.

2 Pose Recovery Using Depth Maps

One of the main advantages of the Kinect™ device is its ability to obtain an aligned representation of RGBD data using a cheap and reliable sensor. Different technologies for capturing depth maps, including the structured light technology of Microsoft® Kinect™, are summarized in Figure 1. The Kinect™ infrared sensor displays a structured/codified matrix of points through the environment. Then, each depth pixel is computed by sampling the derivative of the higher resolution infrared image taken in the infrared camera. This value is inversely proportional to the radius of each gaussian dot, which is linearly proportional to the actual depth. Given the extra image dimension offered by the Kinect™ sensor, new approaches taking benefit of this issue have been proposed for improving PR. The most recent and relevant approaches are described below.

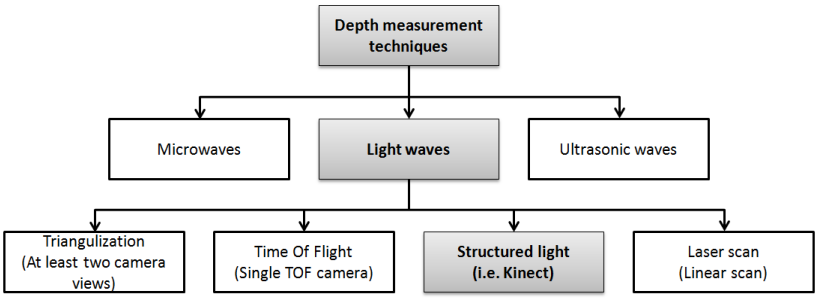

Fig. 1. Different technologies for the acquisition of depth maps

One of the first contributions for PR in depth maps is the approach of [6], which also is part of the core of the Kinect™ device software. The method is based on inferring pixel label probabilities through Random Forest (RF) based on learning offsets of depth features. Then, mean shift is used to estimate human joints and representing the body in skeletal form. An example of the synthetic generated samples for training the system and two trees of the forest are shown in Fig. 2(a). Using the same philosophy, the authors of [24] optimize the results of [6] including a second optimization layer to the RF probabilities in a multi-label Graph Cuts optimization procedure. The scheme and results of this approach are shown in Fig. 2(b). The Graph Cuts theory was previously applied to PR in RGB computer vision approaches [25, 26] as well as other image modalities [27] with successful results.

The skeletal model defined by previous approaches is being used in combination with other techniques in different HBA approaches. For instance, the authors of [28] use the skeletal model in conjunction with computer vision techniques to detect complex poses in situations where there are many interacting actors.

The authors of [29] propose an hybrid approach as an alternative to the Graph Cuts optimization of [24] to static pose estimation, called Connected Poselets. This representation combines aspects of part-based and example-based estimation, first detecting poselets extracted from the training data. The method applies a modified Random Decision Forest to identify Poselet activations. By combining keypoint predictions from poselet activations within a graphical model, the authors infer the marginal distribution over each keypoint without using kinematic constraints. An example of the procedure is illustrated in Fig. 2(c).

In the scope of Probabilistic Graphical Models, different approaches have been proposed. In [33], the authors propose a method for learning shape models enabling accurate articulated human pose estimation from a single image. The authors learn a generative model of limb shape which can capture the wide variation in shape due to varying anatomy and pose. The model is learnt from silhouette, depth, and 3D pose data.

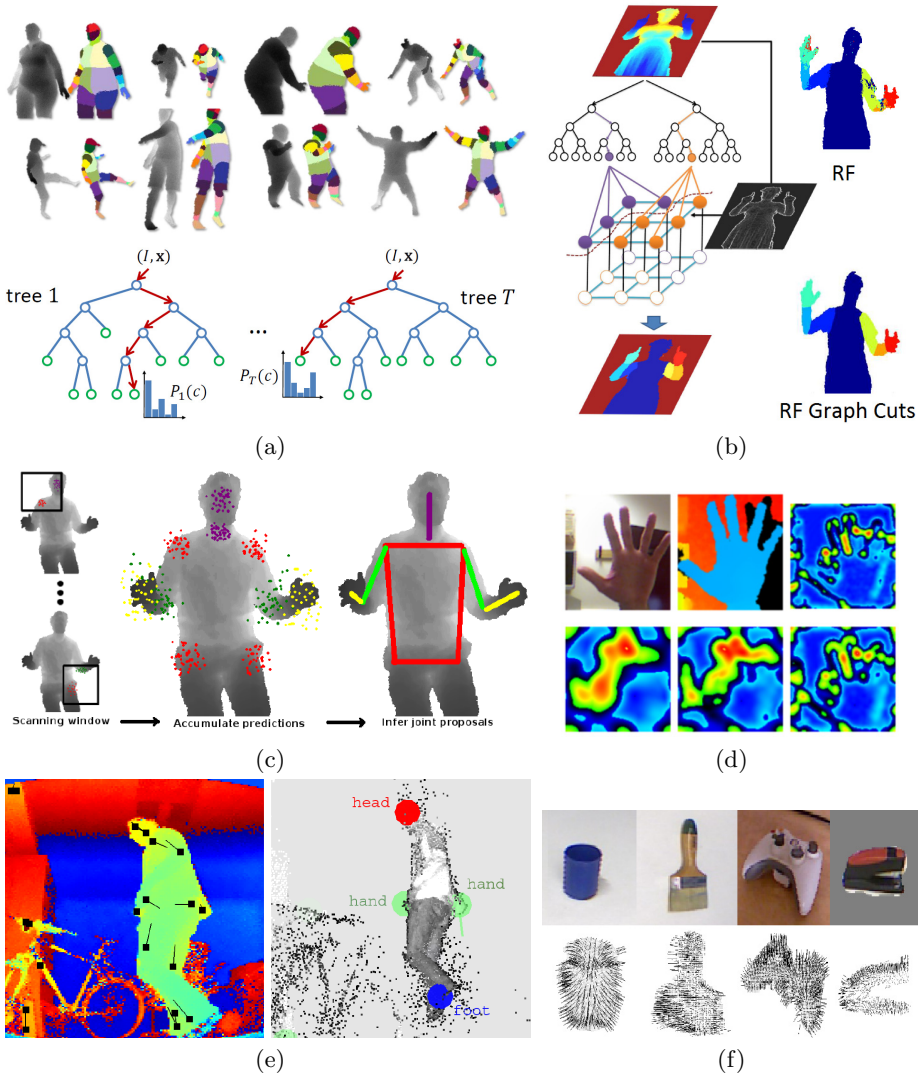

Fig. 2. Different approaches for PR. (a) Random Forest (RF) [6]; (b) Graph Cuts optimization of RF probabilities [24]; (c) Multiscale scanning window. Each window is evaluated by a Random Forest classifier to detect poselet activations [29]. Each activated poselet makes local predictions for the body parts it was trained on. The overall configuration is inferred by combining keypoint predictions within a graphical model, and finding the maximum over the marginal distributions; (d) Gabor filter over depth maps at multiple scales for hand detection in a multi hand pose Random Forest approach for American Sign Language (ASL) recognition [30]; (e) Left: keypoint detection over depth map. Right: inferred body parts from a graphical model [31]; and (f) Normal vectors of segmented objects from depth maps [32].

Obviously, a relevant aspect of the RGBD representation is the definition of new descriptors based on 3D point clouds. On the one hand, recent publications regarding image descriptions are related to the distribution of surface normal vectors [14]. On the other hand, standard computer vision descriptors are used over depth maps instead of RGB data. For instance, the authors of [30] have recently proposed an approach for American Sign Language recognition applying Gabor filters over depth maps of hand regions. Hand-shapes corresponding to letters of the alphabet are characterized using appearance and depth images and classified using Random Forests. An example of the descriptors is shown in 2(d). The authors of [31] propose a novel keypoint detector based on saliency of depth maps which is stable to certain human poses and they include this novel detector in a probabilistic graphical model representation of the human body. The interest points, which are based on identifying geodesic extrema on the surface mesh, can be classified as, e.g., hand, foot, or head using local shape descriptors. This approach also provides a natural way of estimating a 3D orientation vector for a given interest point. This can be used to normalize the local shape descriptors to simplify the classification problem as well as to directly estimate the orientation of the body parts in space. An example of the use of this approach is illustrated in Fig. 2(e). The surveillance system proposed in [32] uses the orientation-invariant Fast Point Feature Histogram [14] based on distribution of normal vectors to identify the robbery of objects in outdoor and indoor environments. An illustration of normal vectors computed from depth maps of image objects is shown in 2(f). In citeKD and [34] the authors use Kernel Descriptors (KD) and Hierarchical Kernel Descriptors (HKD) to avoid the need for pixel attribute discretization, being able to turn any pixel attribute into compact patch-level features. Using KD over multi-modal RGBD, the similarity between two patches is based on a kernel function, called match kernel, that averages over the continuous similarities between all pairs of pixel attributes in the two patches. This method has been recently applied to multiple object recognition in RGBD data with successful results [34]. With a similar idea, the authors of [35] propose the Wave Kernel Signature (WKS) descriptor for 3D keypoint matching, where WKS is represented as the average probability of measuring a quantum mechanical particle at a specific location.

Following the description based on the Geodesic maps from [31], the authors of [36] compute 3D geodesic maps in depth maps for learning distances corresponding to body parts. First, the approach detects anatomical landmarks in the 3D data and, then, a skeleton body model s fitted using constrained inverse kinematics. Instead of relying on appearance-based features for interest point detection, which can vary strongly with illumination and pose changes, the authors build upon a graph-based representation of the depth data that allows the measurement of geodesic distances between body parts. As these distances do not change with body movement, one can localize anatomical landmarks independent of pose. For differentiation of body parts that occlude each other, the authors also use motion information obtained from optical flow. An example of the computed geodesic maps are shown in Fig. 3(a).

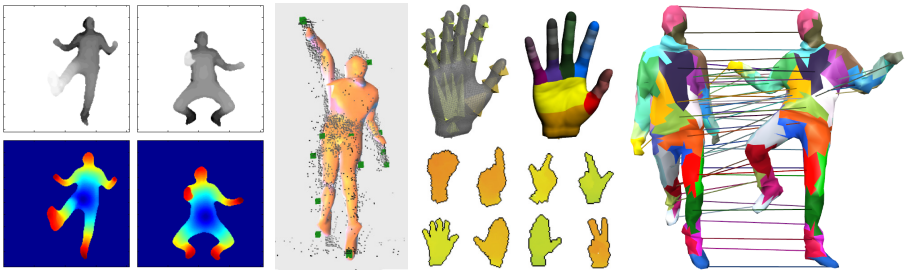

Fig. 3. Different approaches for PR. (a) Geodesic maps over segmented subjects in [36]; (b) Probabilistic graphical model of [37]; (c) Up: Hand labeled model of [38]. Down: Hand Shape Templates of [39]; (d) Linear Programming matching based on kernel descriptors of [40].

Another approach in the scope of graphical models is the probabilistic MOCAP approach of [37]. The authors combine a generative model with a discriminative model that feeds datadriven evidence about body part locations. In each filter iteration, the authors apply a type of local model-based search that exploits the nature of the kinematic chain. As fast movements and occlusion can disrupt the local search, they utilize a set of discriminatively trained patch classifiers to detect body parts. This noisy evidence about body part locations is propagated up the kinematic chain using the unscented transform. The resulting distribution of body configurations allows to reinitialize the model-based search, which in turn allows the system to recover from temporary tracking drift. An example of the graphical model and the inferred pose configuration for a test sample is shown in Fig. 3(b).

Since the description of body posture may vary from each particular application, it is common to find methods that focus on particular limbs to perform a more detailed local description. This is the case of head and hand regions, which described using RGBD data become a very useful tool for different real applications, such as Human Computer Interaction systems (HCI). In the top Fig. 3(c) an example of a hand parts model defined in [38] to train a Random Forest approach is shown. The bottom of Fig. 3(c) shows the shape templates proposed in [39] to look for different hand configurations in a HCI system.

Although most of previous approaches do not require from a previous background extraction step, several methods in literature use to start with an initial background extraction step based on depth information. However, background substraction can lead with several foreground objects from which persons should be identified. In this sense, the authors of [41] present a method for human detection and segmentation based on contouring depth images, applying 2D Chamfer match over silhouettes. Other classical and recent approaches used for shape analysis and matching of point clouds are Active Shape Models, Shape Context, Template Matching, or Linear Programming Approaches [40]. Fig. 3(d) shows an example of body shape matching using the linear programming approach of [40].

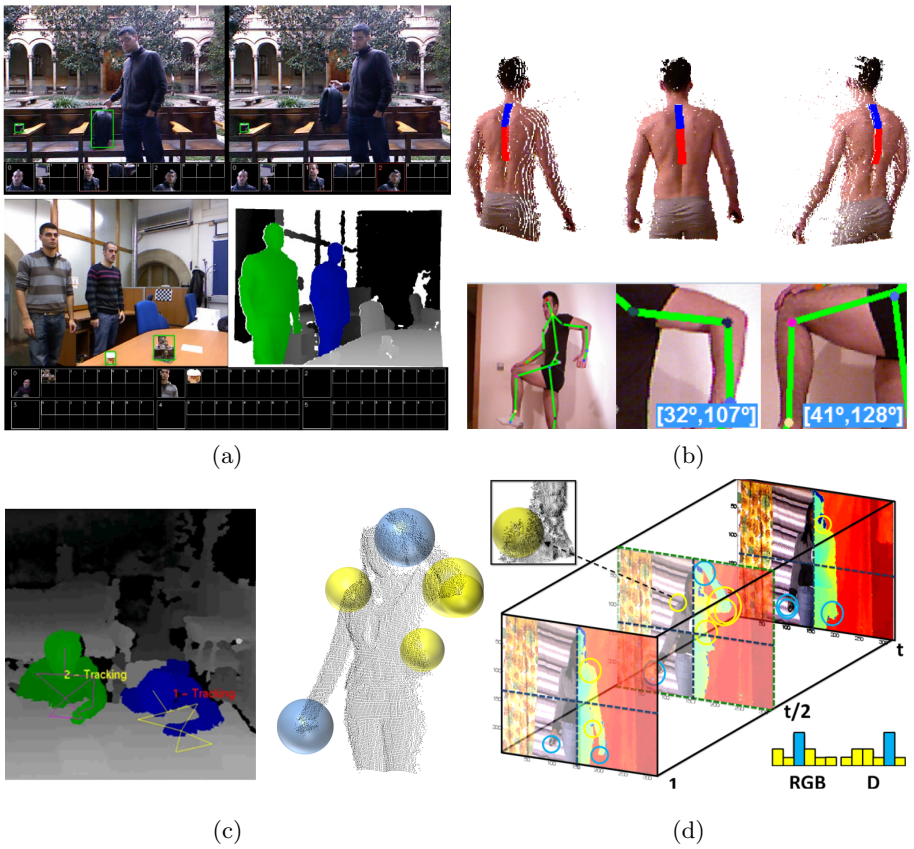

Fig. 4. Some applications of the Human Pose Recovery and Behavior Analysis group (HuPBA) [42]. (a) Multi-modal surveillance system of [32]. Up: Outdoor scenario, user is identified, theft is recognized, and different objects, included a small cup, are detected. Down: Users and objects are correctly identified and classified/recognized. User-object membership relations are defined. Different users can be identified simultaneously by the system; (b) System for static pose and range of movement estimation in physiotherapy, rehabilitation, and fitness condition; (c) Behavior modeling of children with Attention Deficit Hyperactivity Disorder diagnosis; And (d) General modeling of actions and gestures by multi-modal spatio-temporal Bag-of-Visual-and-Depth-Words model. Left: Salient points based on depth maps are detected and labeled to different clusters. Right: Depth descriptors are combined with RGB data descritors, and a Spatio-Temporal Bag-of-Visual-and-Depth-Words for action recognition is defined.

3 Conclusion

In this paper, the recent literature related to Pose Recovery and Human Behavior Analysis from multi-modal RGBD data was reviewed. In particular, the main benefits from the multi-modal RGBD from Microsoft® Kinect™ were described. Among the broad range of applications related to the devices, those particular methodologies related to PR for improved Human Behavior Applications were discussed.

We saw recent approaches based on background substraction and body part models detection and segmentation, using both discriminative and generative approaches, including hybrid approaches. Moreover, classical descriptors and techniques from 2D computer vision methodologies have been redefined and extended using the extra dimension provided by the Kinect™ device. This has lead to the design of new descriptors, use of normal vectors of surfaces, geodesic paths in 3D spaces, clustering of point clouds, etc. As a result, several non-invasive applications have become real. In Fig. 4 some applications of our research group using RGBD data representation are briefly described [42].

Besides the high performance of current methods for PR and HBA using RGBD data, several issues remain opened and require further attention. For instance, background extraction in complex scenarios still becomes difficult when the foreground object is close to artifacts belonging to the background. In those cases, the use of depth information is not straightforward, and more sophisticated approaches are required. In the case of PR, though one can benefit from the high discriminative power of RGBD representation, more accurate recognition of poses under different appearance and points of view are also necessary for some real applications. This requires dealing with the whole set of human body deformations, including occlusions and changes in appearance produced, for example, by clothes and non-controlled environmental factors. Furthermore, some real applications also need higher resolution of depth maps to be applied in real environments and deal with the reflectance deviations of the infrared light for particular materials. On the other hand, significant advances are expected to appear in the next years regarding the hardware capabilities.

References

1. Jain, H., Subramanian, A.: Real-time upper-body human pose estimation using a depth camera, HP Technical Reports
2. Rodgers, J., Anguelov, D., Hoi-Cheung, P.: Object pose detection in range scan data. In: CVPR, pp. 2445–2452 (2006)
3. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera. In: CVPR, pp. 755–762 (2010)
4. Sabata, B., Arman, F., Aggarwal, J.: Segmentation of 3d range images using pyramidal data structures. CVGIP: Image Understanding 57(3), 373–387 (1993)
5. Primesensor™, `http://www.primesense.com/?p=514`
6. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M.: Real-time human pose recognition in parts from single depth images (2011)
7. Dephsense ds311, `http://www.softkinetic.com/Solutions/DepthSensecameras.aspx`

8. Openni, http://www.openni.org
9. Flexible action and articulated skeleton toolkit (faast), http://projects.ict.usc.edu/mxr/faast/
10. Suma, E., Lange, B., Rizzo, A., Krum, D.M.: FAAST: the flexible action and articulated skeleton toolkit. In: Virtual Reality, Singapore, pp. 245–246 (2011)
11. Kinect for windows sdk from microsoft research, http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/
12. Openkinect (libfreenect), http://openkinect.org/
13. Code laboratories cl nui platform - kinect driver/sdk, http://codelaboratories.com/nui/
14. Point cloud library (pcl), http://pointclouds.org/
15. Rusu, R.B.: Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. Articial Intelligence (KI-Kuenstliche Intelligenz) (2010)
16. Lai, K., Bo, L., Ren, X., Fox, D.: Sparse distance learning for object recognition combining rgb and depth information. In: ICRA
17. Bo, L., Ren, X., Fox, D.: Depth kernel descriptors for object recognition. In: IROS, pp. 821–826 (2011)
18. Koch, R., Schiller, I., Bartczak, B., Kellner, F., Koser, K.: Mixin3d: 3d mixed reality with tof-camera, pp. 126–141 (2009)
19. Castaneda, V., Mateus, D., Navab, N.: Slam combining tof and high-resolution cameras. In: WACV, pp. 672–678 (2011)
20. Gehrig, D., Kuehne, H.: Hmm-based human motion recognition with optical flow data. In: IEEE International Conference on Humanoid Robots, Humanoids 2009 (2009)
21. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. CVIU 104(2-3), 210–220 (2006)
22. Zhou, F., la Torre, F.D., Hodgins, J.K.: Aligned cluster analysis for temporal segmentation of human motion. In: IEEE Conference on Automatic Face and Gestures Recognition, FG (2008)
23. Reyes, M., Dominguez, G., Escalera, S.: Feature weighting in dynamic time warping for gesture recognition in depth data. In: ICCV, Barcelona, Spain (2011)
24. Hernandez-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., Escalera, S.: Graph cuts optimization for multi-limb human segmentation in depth maps. In: CVPR (2012)
25. Hernandez-Vela, A., Reyes, M., Escalera, S., Radeva, P.: Spatio-temporal grabcut human segmentation for face and pose recovery. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, CVPR (2010)
26. Hernandez-Vela, A., Primo, C., Escalera, S.: Automatic user interaction correction via multi-label graph cuts. In: 1st IEEE International Workshop on Human Interaction in Computer Vision HICV, ICCV (2011)
27. Igual, L., Soliva, J., Hernandez-Vela, A., Escalera, S., Jimenez, X., Vilarroya, O., Radeva, P.: A fully-automatic caudate nucleus segmentation of brain mri: Application in volumetric analysis of pediatric attention-deficit/hyperactivity disorder. In: BioMedical Engineering OnLine (2011)
28. Liu, Y., Stoll, C., Gall, J., Seidel, H.: Markerless motion capture of interacting characters using multi-view image segmentation. CVPR 14(1), 1249–1256 (2011)
29. Holt, B., Ong, E.-J., Cooper, H., Bowden, R.: Putting the pieces together: Connected poselets for human pose estimation. In: ICCV (2011)
30. Pugeault, N., Bowden, R.: Spelling it out: Real-time asl fingerspelling recognition. In: ICCV (2011)

31. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: ICCV, pp. 3108–3113 (2011)
32. Clapes, A., Reyes, M., Escalera, S.: User identification and object recognition in clutter scenes based on rgb-depth analysis. In: Articulated Motion of Deformable Objects (2012)
33. Charles, J., Everingham, M.: Learning shape models for monocular human pose estimation from the microsoft xbox kinect. In: ICCV, pp. 1202–1208 (2011)
34. Bo, L., Lai, K., Ren, X., Fox, D.: Object recognition with hierarchical kernel descriptors. In: CVPR (2011)
35. Aubry, M., Schlickewei, U., Cremers, D.: The wave kernel signature: A quantum mechanical approach to shape analysis. In: ICCV (2011)
36. Schwarz, L., 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 (2011)
37. Ganapathiand, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera. In: CVPR, pp. 755–762 (2010)
38. Keskin, C., Racc, F., Kara, Y., Akarun, L.: Real time hand pose estimation using depth sensors. In: ICCV (2011)
39. Minnen, D., Zafrulla, Z.: Towards robust cross-user hand tracking and shape recognition. In: ICCV, pp. 1235–1241 (2011)
40. Windheuser, T., Schlickewei, U., Schmidt, F.R.: Geometrically consistent elastic matching of 3d shapes: A linear programming solution. In: ICCV (2011)
41. Xia, L., Chen, C.-C., Aggarwal, J.K.: Human detection using depth information by kinect department of electrical and computer engineering. PR, 15–22 (2011)
42. Human pose recovery and behavior analysis group, `http://www.maia.ub.es/~sergio/`