Automatic Hand Detection in RGB-Depth Data Sequences

Vitaliy KONOVALOV [a], Albert CLAPÉS [a,b] and Sergio ESCALERA [a,b]
[a] *Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain.*
[b] *Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Spain.*

Abstract. Detecting hands in multi-modal RGB-Depth visual data has become a challenging Computer Vision problem with several applications of interest. This task involves dealing with changes in illumination, viewpoint variations, the articulated nature of the human body, the high flexibility of the wrist articulation, and the deformability of the hand itself. In this work, we propose an accurate and efficient automatic hand detection scheme to be applied in Human-Computer Interaction (HCI) applications in which the user is seated at the desk and, thus, only the upper body is visible. Our main hypothesis is that hand landmarks remain at a nearly constant geodesic distance from an automatically located anatomical reference point. In a given frame, the human body is segmented first in the depth image. Then, a graph representation of the body is built in which the geodesic paths are computed from the reference point. The dense optical flow vectors on the corresponding RGB image are used to reduce ambiguities of the geodesic paths' connectivity, allowing to eliminate false edges interconnecting different body parts. Finally, we are able to detect the position of both hands based on invariant geodesic distances and optical flow within the body region, without involving costly learning procedures.

Keywords. Human-Computer Interaction, Hand detection, Human Pose Recovery, Optical Flow, Geodesic paths, Multi-modal RGB-Depth data

1. Introduction

Detecting hands in multi-modal visual data has become a challenging Computer Vision problem with several applications of interest. Particularly, detecting hands from visual RGB and depth data is an specially hard task because of several difficulties we need to tackle, including changes in illumination, viewpoint variations, the articulated nature of the human body and particularly the high flexibility of the wrist articulation, and also the deformability of the hand itself.

Automatic hand detection from visual data can be seen as a specification of the more general problem of human pose recovery. As many other problems in Computer Vision, the one of human pose recovery can be divided in two different fashions: learning-based approaches consisting in learning beforehand from data [2,3,8,15] and, in the other hand, those based in body parameter estimation from observed features, without introducing an initial learning step [4,9,5,10,18]. Regarding the first, the work of Jaeggli et al. [3] predict the pose of 2D silhouettes with a trained pose estimator in walking/running activities. In [8], the pose prediction is performed in 3D voxel-structured data, but also for pre-determined activities. A disadvantage of the learning-based strategy is that it is limited to the amount of training data and the casuistics provided to the learning algorithm. There have been proposed methods which do not use prior knowledge, but that are highly dependent on the reliability of the feature extraction stage as a counterpart. In those approaches, efficient state inference techniques are required to cope with the high dimensionality of the human body and, specially, the enormous hand configuration space. In [4], Kehl and Van Gool tackle the problem of pose recovery setting multiple cameras and generating from their acquisitions precise reconstructions of 3D human volumes. Other authors, as those in [10], thought about assisting the computer vision-related techniques with other complementary technologies, as inertial sensors, in the pose estimation task.

In order to overcome the limitations caused by the acquisition of visual data captured with monocular cameras or the complexity of multi-camera setups, Time-of-Flight (ToF) cameras have been considered by several authors for analysis of human motions [1,11]. These cameras, which are capable to provide the depth value of a pixel (i.e., the actual distance to the camera), became popular because of the huge amount of possibilities they offer at a relative low price. In this sense, Microsoft® has launched Kinect™ [14] — an even cheaper multi-sensory device than ToF cameras — based on structured light technology, which is capable of capturing visual depth information to then generate real-time dense depth maps containing discrete range measurements of the physical scene. The device is so compact and portable that it can be easily installed in any environment to analyze scenarios where humans are present. Many researchers have obtained their first results in the field of human motion capture using this technology. In particular, Shotton et al. [15] presented one of the greatest advances in the extraction of the human body pose from depth maps, which also forms the core of the Kinect™ human recognition framework. More recent works have based on the previous proposal to improve human pose recovery. The authors of [20] improve Random Forest pose estimation using a posterior Graph Cuts optimization procedure. This kind of multi-modal human pose representations have shown to improve standard gesture recognition approaches for HCI systems [17]. In another important work, Plagemann et al. [7] propose to detect and identify some body parts (head, hands, and feet) in depth images by detecting interest points, which are based on identifying geodesic extrema, together with their orientation, and then classify the body parts using a local shape descriptor normalized by the orientation. A later improvement of this work embedded the pose estimation method in a Bayesian tracking framework [6]. Optical flow also has been shown to be useful in segmenting humans as moving objects when a continuous sequence of frames is available. In [12], the optical flow is used to improve the background subtraction in a monocular pedestrian tracking application. In our work, the optical flow is added as a frontier constraint in the minimum-distance geodesic paths computation. Okada et al. [13] proposed a walking tracking application in a stereo setup, in which the disparity (depth information) and the optical flow are combined to estimate the target region state (3-D position, posture, and motion) through the tracking sequence. See [16] for a more detailed review of recent state-of-the-art related approaches.

Inspired by the approach presented in [18], we propose a fully-automatic system for hand detection in multi-modal RGB-Depth data. Our proposed automatic hand detection scheme consists of several steps: (a) the subject's upper body and the desk's surface are segmented from the background by means of Otsu's thresholding, (b) the point cloud representing the foreground is built and downsampled, (c) the planar region in the foreground representing the table is modeled and segmented to come out only with the point cloud representing the subject, (d) an invariant torso reference point is estimated by means of a distance map computed from a 2D projection of segmented subject, and finally (e) the geodesic distances from the reference point to the other points are calculated. For the later task, the downsampled human body is designed as a graph, and minimum path algorithms are applied. Moreover, dense optical flow estimation is used to restrict inter-bodypart connectivity on the graph, yielding an accurate detection of hand regions. The proposed scheme does not require costly training procedures. Furthermore, contrary to [18], our approach does not require neither from any specific initial calibration pose, but instead a geodesic histogram is used to recognize human limbs at near stable geodesic distances, obtaining a fully-automatic and robust detection.

The rest of the paper is organized as follows. Section 2 presents the system for automatic hand detection from multi-modal RGB-Depth data sequences. Section 3 presents the evaluation of the proposed system on real data from different scenarios, and finally, Section 4 concludes the paper.

2. Method

Our aim is to settle an automatic hand detection framework in multi-modal RGB-Depth data, accurate and robust enough to be applied in many HCI applications. In this work, we restrict the procedure to frontal upper body scenarios, though the system is general to be applied to other user-case scenarios and to detect different body limbs. Similar to the approach presented in [18], our method is based on the assumption that all anatomical landmarks (hands in our case) remain at a nearly constant geodesic distance for an estimated anatomical reference point. However, differently from the approach of [18], our system does not require any calibration pose for initialization. The different steps of the system are shown in Figure 1. Next, we describe each stage of the system in detail.

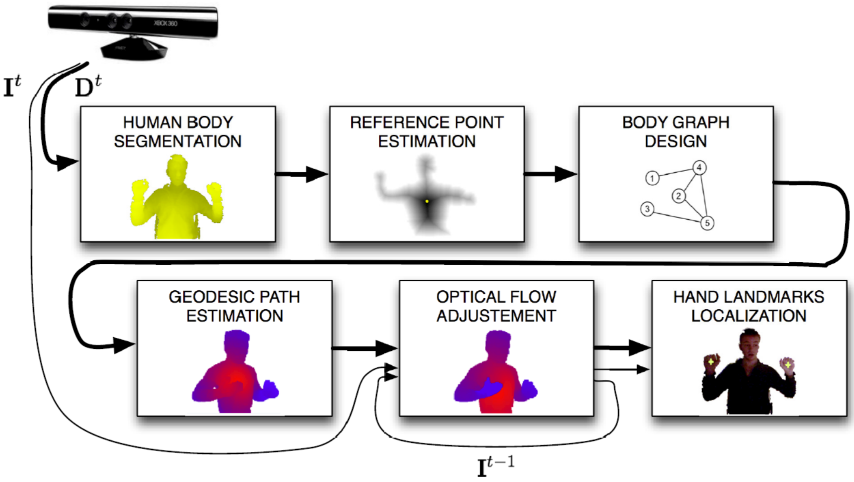


Figure 1. Pipeline of the system for automatic hand detection in multi-modal RGB-Depth data sequences.

2.1. Human body segmentation

Given a multi-modal RGB-Depth data sequence at the current time instant t, the first step consists in segmenting the user from the rest of the scene. For this task, and given the nature of our scenarios, we perform a foreground segmentation in the depth frame $\mathbf{D}^t$ based on the depth value in each range pixel. Thus, assuming a bimodal depth distribution in the depth image $\mathbf{D}^t$, we apply Otsu's method to automatically find the appropriate threshold value $\alpha(\mathbf{D}^t)$, the one that better separates the two modalities corresponding to the foreground objects and the background. In our case, foreground objects are the user together with the desk's planar surface. Thus, a second step is needed to separate the subject's body from the desk's surface if it is present in the image.

The segmented foreground is transformed to a point cloud (using the intrinsic Kinect™ parameters) and downsampled using a voxel grid filter. This filtering step consists in partitioning the point cloud space as a voxel grid of s-sized voxels (or grid cells), and in each voxel keeping just the centroid of all the originally contained points. Let $\mathbf{P}^t = \{\mathbf{p}_{ijk}\}$ denote the filtered point cloud representing the foreground segmented region. The notation indicates that the point $\mathbf{p}_{ijk}$ corresponds to the filtered point in the grid cell (i, j, k). This downsampling step will greatly speed up the subsequent stages.

Finally, in order to remove the desk point cloud part from the segmented foreground, we estimate the biggest planar region [1] in $\mathbf{P}^t$, and remove it. This simple approach achieves very robust results removing such non-subject artifact from the point cloud. In case of facing more complex environments as, for instance, a clutter desk with different non-planar and non-static objects on it, more cumbersome foreground segmentation approaches should be applied instead. The remaining points $\mathbf{B}^t$ will be the ones considered for designing the human graph and estimating posteriorly geodesic distances for hand detection.

2.2. Body graph design

It is constructed a graph $G^t = (V^t, E^t)$, where $V^t = \mathbf{B}^t$ are the vertices and $E^t \subseteq V^t \times V^t$ are the edges. Two vertices are connected in the graph with an edge based on their vicinity in the 3D point cloud. The set of edges is defined as:

$$E^t = \{(\mathbf{p}_{ijk}, \mathbf{p}_{i'j'k'}) \in V^t \times V^t : \quad \| (i,j,k)^T - (i',j',k')^T \|_\infty < 1\}, \tag{1}$$

where $\| \, . \, \|_\infty$ is the maximum norm and $(i,j,k)^T - (i',j',k')^T$ are the 3D coordinates of a pair of points $\mathbf{p}_{ijk}$ and $\mathbf{p}_{i'j'k'}$ in $\mathbf{B}^t$. In other words, two points $\mathbf{p}$ and $\mathbf{p}'$ are connected by a graph edge if they are neighbors in a 3D neighborhood of 27 voxels. Moreover, for each edge $e = (\mathbf{p}, \mathbf{p}') \in E^t$, we store a weight $w(e) = \| \mathbf{p} - \mathbf{p}' \|_2$, used later in the geodesic paths computation.

2.3. Reference point estimation to compute geodesic paths

After having defined the body graph G^t from the subject's point cloud $\mathbf{B}^t$, next we re-process this point cloud in order to detect a reference point to start the geodesic map computation. The reference landmark $\mathbf{p}^t_{ref} \in \mathbf{B}^t$ corresponds to that torso point that has

[1] `http://pointclouds.org/documentation/tutorials/planar_segmentation.php`

its corresponding planar projection to a pixel $\mathbf{x}^t_{ref}$ that is at maximum distance to all the contour pixels of the projected subject's silhouette. A torso pixel estimation is graphically shown in Figure 2(a). We found this reference point more stable in our dataset scenario than other common points, such as head or neck.

In order to compute the reference torso point $\mathbf{p}^t_{ref}$, we project the segmented human body point cloud $\mathbf{B}^t$ into a 2D image and compute external contour. This contour map C corresponds to the external silhouette of the body and it is not affected by internal body contours. The contour needs to be processed by means of mathematical morphology, so as to obtain a closed region for its further reliable use. Then, each point within the silhouette in the 2D contour image is assigned the minimum Euclidean distance to the closest point to the contour map C, and we assign to the reference torso pixel the one with the highest value from the computed distance map:

$$\mathbf{x}^t_{ref} = \mathsf{argmax}_{\mathbf{x}} \mathsf{min}_{\mathbf{x}_C \in C} d(\mathbf{x}, \mathbf{x}_C)$$

where $\mathbf{x}$ take the values of the pixels inside the silhouette, $\mathbf{x}_C$ the contour pixels, and $\mathbf{x}^t_{ref}$ is the reference torso pixel. This computation can be efficiently computed in linear time. Then our torso reference pixel $\mathbf{x}^t_{ref}$ is re-projected to the 3D point cloud in order to compute the geodesic map starting at that point $\mathbf{p}^t_{ref}$.

2.4. Geodesic path estimation

Using G^t, we are able to measure geodesic distances between different body locations. The geodesic distance $d_G(\mathbf{p}, \mathbf{p}')$ between two points $\mathbf{p}, \mathbf{p}' \in V^t$ is given by:

$$d_G(\mathbf{p}, \mathbf{p}') = \sum_{e \in EP(\mathbf{p}, \mathbf{p}')} w(e),$$

where $EP(\mathbf{p}, \mathbf{p}')$ contains all edges along the shortest path between $\mathbf{p}$ and $\mathbf{p}'$ using min-path Dijkstra's algorithm. Intuitively, the geodesic distance between two locations of the body is the length of the shortest path over all the body surface. Eventually, we would be able to perform the hand detection. However, the accuracy of the detection increases including some restrictions in the graph nodes connectivity based on estimated motion information.

2.5. Including restrictions based on optical flow

The high deformability of the human body, and in particular the upper limbs, leads to challenging disambiguation problems. Given the more straightforward case of having the arms enough separated from each other and also from the human torso, we can detect both hand landmarks directly with the procedure described up to this point. However, in cases where the arms are stick together to another part of the body, i.e. having at least two points $\mathbf{p}, \mathbf{p}' \in V^t$ that belong to two different body parts satisfying the condition 1 and hence establishing an edge $(\mathbf{p}, \mathbf{p}') \in E^t$, we may have undesired graph connections between those parts, and thus an incorrect estimation of the geodesic paths. Figure 2(e) gives an example where an arm in front of the torso is connected to the upper body and the geodesic distance from the body center to the hand is underestimated. Without correction, hand landmarks cannot be detected properly in some cases. We therefore propose a disambiguation approach, similar to the one proposed in [18], that makes use of movement occurring between each pair of frames. Assuming that distinct body parts move separately, this approach allows us to disconnect points (removing the graph edge) belonging to different body parts.

At each time step, the dense optical flow vectors are computed between the pair of intensity images I^{t-1} and I^t, which are grayscaled RGB images that have been previously aligned and synchronized with the depth frames in the same instant of their acquisition. From this computation, we obtain $\mathscr{F}^t = (\mathbf{F}^t_x, \mathbf{F}^t_y)$, being $\mathbf{F}^t_x(i,j)$ and $\mathbf{F}^t_y(i,j)$ the movement of an intensity pixel (i,j) between the two images in the x and y-direction respectively. Figure 2(b) shows an example of an estimated optical flow map, remarking the pixels containing the highest magnitude of the computed optical flow between consecutive frames. Using the updated and corrected map, we can remove the undesired edges in the graph G^t that connect points on overlapping body parts. Therefore, the set of edges are updated as $E^t = E^t - F^t$, with:

$$F^t = \{(\mathbf{x}_{ij}, \mathbf{x}_{kl}) \in E^t : \quad \| \, |\mathbf{F}^t(i,j)| - |\mathbf{F}^t(k,l)| \, \|_2 > \beta\},$$

where β is a threshold value of optical flow gradient magnitude and $|.|$ contains the magnitude of a vector. The magnitude of an optical flow vector at pixel location (i,j) at instant t is:

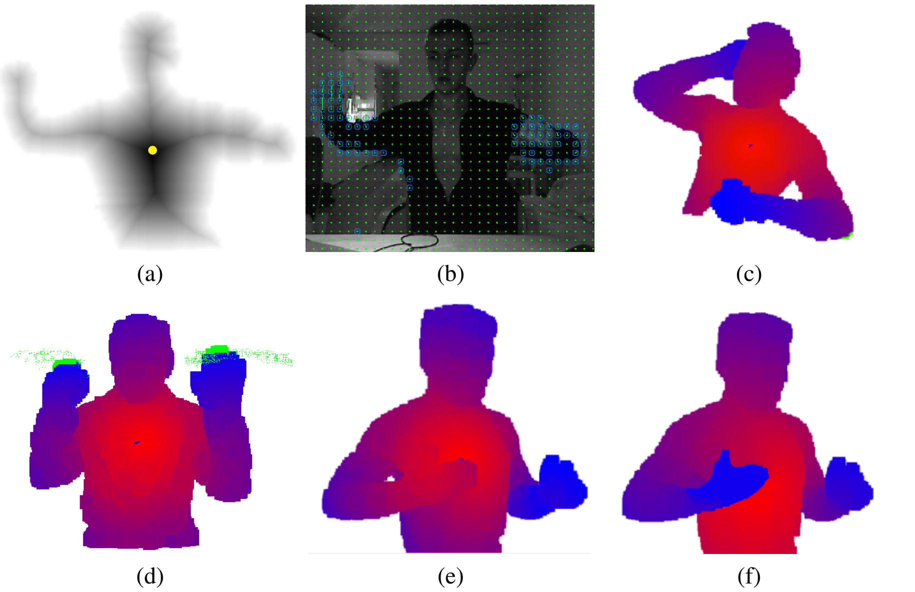
$$|\mathbf{F}^t(i,j)| = \sqrt{\mathbf{F}^t_x(i,j)^2 + \mathbf{F}^t_y(i,j)^2}$$

Figure 2(f) shows the example of Figure 2(e) removing graph connections based on highest values of the optical flow. Note that in this case, the geodesic map is successfully estimated, obtaining similar geodesic values for both user hands.

2.6. Automatic detection of hands

Once the geodesic map has been computed started at point $\mathbf{p}^t_{ref}$ based on the constructed graph and considering optical flow restrictions on edges, we construct a histogram of geodesic values H_G in order to look for the range of geodesic distances that codify the geometrical constraints of each body limb. The histogram H_G encodes the distribution of geodesic distances $d_G(\mathbf{p}, \mathbf{p}')$ for all active nodes in the graph. An example of the histogram H_G encoding the distribution of geodesic distances is shown in Figure 3. In our scenario for automatic hand detection in HCI environments, since upper body regions are the ones computed by geodesic paths starting at the center torso point, the highest values of H_G correspond to the hand regions (for both hands), and in the way that we reduce those values, we can find head and inner torso points at the lowest values of H_G. In this sense, we experimentally found that keeping the 1% highest values of the histogram we capture most hand regions while preventing the detection of false positive regions. An example of the final hand voxel detection is shown in Figure 2(d).

Furthermore, in order to prevent some noisy detections, after detecting hand points, we filter the segmented points by simple mathematical morphology operators, keeping the two highest connected components and saving the mean center of coordinates for each one. In addition, temporal coherence is taken into account to verify the detection of a hand region in a lattice around previous detection by a distance threshold defined by γ. This also allows us to do not detect hands if they are occluded and ensures smoothness in the detected hand trajectories. In Figure 2(d), the green cloud around the hands corresponds to their detections in consecutive frames. Note the good definition of hand trajectory using the proposed system for this particular multi-modal frames sequence.


(a) (b) (c)

(d) (e) (f)

Figure 2. (a) Distance map computation from a foreground segmented subject. (b) Optical flow estimation. (c) Example of geodesic map computation from an initial torso point. (d) Trajectory of hand detection (detected points in green) from highest values of geodesic maps in an image sequence. (e) Geodesic map computation without optical flow. (f) The same data applying optical flow to disambiguate inter-bodypart graph connectivity.

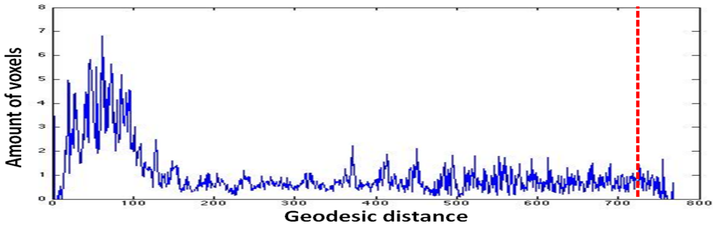


Figure 3. Geodesic histogram H_G encoding the geodesic distances' distribution for an input processed image. The red dotted line corresponds to the threshold to segment both hand pixels/points.

3. Results

In order to present the results, first we describe the data, settings, and validation measurements of the experiments.

Data: We recorded a data set consisting of 6 different users performing different natural uncontrolled gestures in front of a Kinect$^{\text{TM}}$ device simulating HCI scenarios. Upper body multi-modal sequences containing RGB and depth data were recorded. The data set has a total of 3000 RGB plus their corresponding 3000 depth frames of resolution 640×480 pixels. In this data set, hand locations (when visible) were labeled using 3D world coordinates, representing a total of 2171 annotated hands. Some examples of the data set are shown in Figure 4.

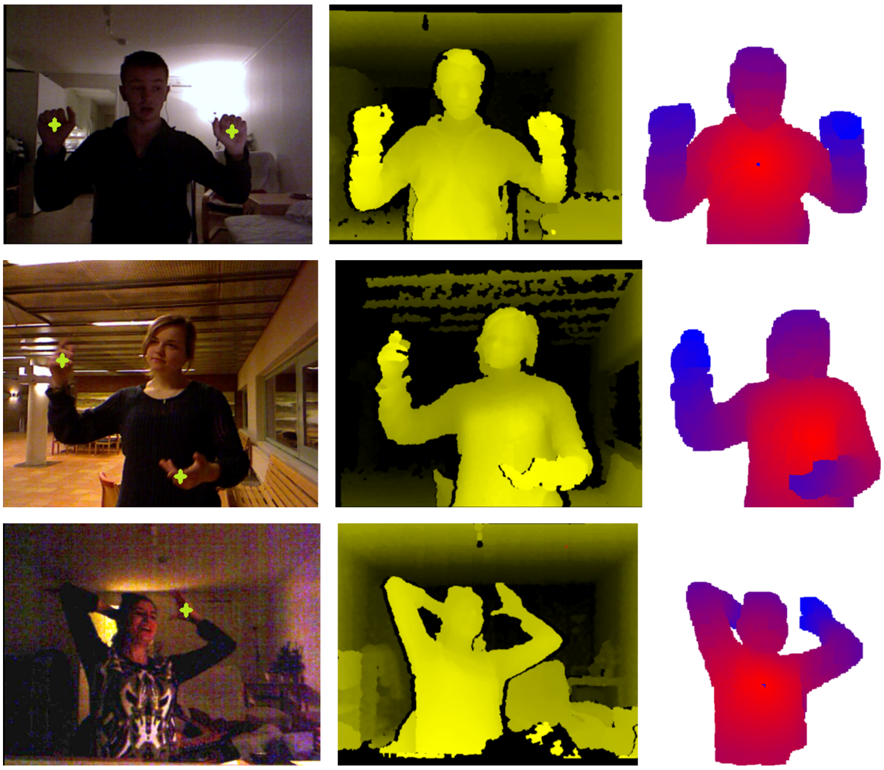


Figure 4. Examples of the data set. From left to right: RGB image and annotated hand locations, aligned depth data, and computed geodesic map applying the proposed procedure.

Settings: For all the experiments, we experimentally set the parameters $\delta = \sqrt{2}$ mm, $\beta = 0.2$, and $\gamma = 5$ mm, respectively.

Validation measurements: We compute the number of detected hands by means of our approach, based on the ground truth labels for each image hand and different tolerance values of a threshold in 3D real coordinates. The final performance is shown as the percentage of correctly detected hands.

3.1. Experiments

In order to test the accuracy of our automatic hand detection approach, we tested our proposal on the designed data set. To compare groundtruth with estimated hand locations, from the automatically detected hand regions, we estimate the center of mass of each connected component, which is compared with the groundtruth data. Looking for the number of correctly detected hands, we introduce a tolerance distance threshold parameter λ expressed in millimeters and test for the recognition accuracy for different values of λ. For our experiments, we fixed a maximum value of λ to 25 mm just to estimate accurate detections useful for precise human computer interaction systems. We also test for different percentage of highest geodesic points within H_G to look for the sensitivity and best values for this parameter. We test the accuracy just applying geodesic computation from the initial detected center torso point and also including the optical flow restriction with $\beta = 0.2$. The results are shown in Figure 5. In Figure 5(a), one can see that the best performance is achieved for 25 mm of tolerance distance and 1% of maximum geodesic distances from the mean geodesic map histogram H_G. As it is expected, if we increase the distance tolerance better accuracy will be achieved. However, we only want to consider precise detection in order to allow the system for accurate HCI requirements. When increasing the percentage of geodesic values, a larger number of hand points are detected, and in consequence, the center of mass of the detected region is displaced in relation to the ground truth annotation, resulting in a reduction of the detected hands. In Figure 5(b) one can see the same results including optimal flow restrictions within the geodesic map estimation. As it is shown, the best performance is achieved for similar values of the method parameters. In this case, the accuracy is increased in a range between 5%-10% in relation to the results provided in Figure 5(a). Best results for both strategies are numerically shown in Table 1.

The main cases where our approach fails are mainly because of two reasons. First, though we are able to detect hands, not always the 1% of maximum geodesic points is the optimal number for a subset of images. In consequence the mean hand point is displaced, without satisfying the maximum spatial distance restriction of 25 mm. And second, some arm/hand configurations in front the device occlude surface information about the connectivity of the regions. As a result, geodesic path can not connect different parts of the arm, and hand are lost in that case. One possible solution in to extend the system including an extra calibrated device, reconstructing the human point cloud to allow connecting all the surface points, reducing the percentage of occluded points.

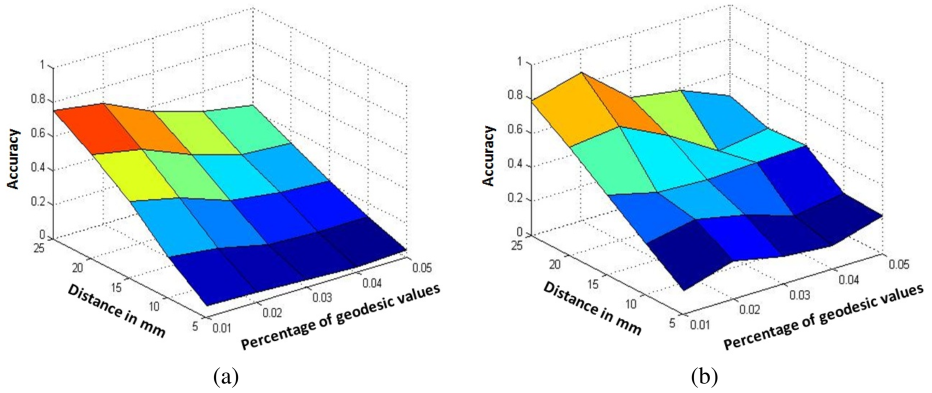


Figure 5. (a) Classification accuracy of automatic hand detections applying geodesic path estimation without optical flow restrictions. (b) Classification accuracy of automatic hand detections applying geodesic path estimation with optical flow restrictions.

	Geodesic path	Geodesic path with optical flow
Classification accuracy	74.12%	**84.15%**

Table 1. Best accuracy of hand detection with geodesic path without/with optical flow.

4. Conclusion

We proposed a system for automatic hand detection in multi-modal data sequences. The method is based on segmentation and graph design of the human body from depth maps. Then, optical flow from RGB data is used to remove edge ambiguities in the graph, and geodesic paths are computed to obtain the geodesic distances from an estimated torso reference point to all the other points conforming the body surface. Geodesic distance values are then used to automatically detect both hand locations. The approach is simple, robust, efficient, and fully-automatic, without requiring a training phase or fixed pose initialization protocols. The results on real multi-modal RGB-Depth data from different ambient conditions and arbitrary subject behaviors show the high accuracy and suitability of the proposal to be applied in real HCI scenarios.

References

[1] *S. Soutschek, J. Penne, J. Hornegger, and J. Kornhuber, 3-d gesture based scene navigation in medical imaging applications using time-of-flight cameras, CVPR Workshops, 2008.*

[2] *R. Urtasun and T. Darrell, Sparse probabilistic regression for activity independent human pose inference, CVPR, 2008.*

[3] *T. Jaeggli, E. Koller-Meier, and L. V. Gool, Learning generative models for multi-activity body pose estimation, IJCV, vol. 83, no. 2, pp. 121-134, 2009.*

[4] *R. Kehl and L. Gool, Markerless tracking of complex human motions from multiple views, CVIU, 2006.*

[5] *J. Bandouch, F. Engstler, and M. Beetz, Accurate human motion capture using an ergonomics-based anthropometric human model, AMDO, 2008.*

[6] *V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, Real time motion capture using a single time-of-flight camera, CVPR, 2010.*

[7] *C. Plagemann, V. Ganapathi, and D. Koller, Real-time identification and localization of body parts from depth images, ICRA, 2010.*

[8] *Y. Sun, M. Bray, A. Thayananthan, B. Yuan, and P. Torr, Regression based human motion capture from voxel data, BMVC, 2006.*

[9] *Y. Zhu, B. Dariush, and K. Fujimura, Controlled human pose estimation from depth image streams, CVPR Workshops, 2008.*

[10] *G. Pons-Moll, A. Baak, T. Helten, M. Muller, H.-P. Seidel, and B. Rosenhahn, Multisensor-fusion for 3d full-body human motion capture, CVPR, pp. 1-8, 2010.*

[11] *R. Jensen, R. Paulsen, and R. Larsen, Analyzing gait using a timeof- flight camera, Scandinavian Conference on Image Analysis, pp. 21-30, 2009.*

[12] *S. Denman, V. Chandran, and S. Sridharan, An adaptive optical flow technique for person tracking systems, PRL, vol. 28, no. 10, pp. 1232-1239, 2007.*

[13] *R. Okada, Y. Shirai, and J. Miura, Tracking a person with 3-d motion by integrating optical flow and depth, FG, pp. 1-6, 2000.*

[14] *Microsoft® Kinect^{TM} for Windows SDK beta programming guide beta 1 draft version 1.1. 2012.*

[15] *J. Shotton, A. W. Fitzgibbon, M. Cook, and T. Sharp, Real-time human pose recognition in parts from single depth images, CVPR, pp. 1297-1304, 2011.*

[16] *Sergio Escalera, Human Behavior Analysis from Depth Maps, AMDO, pp. 282-292, 2012.*

[17] *Miguel Reyes, Gabriel Domnguez, and Sergio Escalera, Feature Weighting in Dynamic Time Warping for Gesture Recognition in Depth Data, 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, ICCV, 2011.*

[18] *L.A. Schwarz, A. Mkhitaryan, D. Mateus, N. Navab, Estimating human 3D pose from Time-of-Flight images based on geodesic distances and optical flow, FG, pp 700-706, 2011.*

[19] *C. Plagemann, V. Ganapathi, and D. Koller, Real-time identification and localization of body parts from depth images, ICRA, 2010.*

[20] *Antonio Hernndez-Vela, Nadezhda Zlateva, Alexander Marinov, Miguel Reyes, Petia Radeva, Dimo Dimov, and Sergio Escalera, Human Limb Segmentation in Depth Maps based on Spatio-Temporal Graph Cuts Optimization, JAISE, 2012.*