DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

Human detection from still depth images

Gulsum Nurdan Can, Helin Dutagaci; Department of Electrical-Electronics Engineering; Eskisehir Osmangazi University; Turkey

Abstract

Human detection from depth images is gaining substantial attention since depth information facilitates object extraction from the background. In this paper, we propose a human detection method where search for humans is performed over regions obtained from a pre-segmentation of the depth image. Our segmentation scheme is based on K-means clustering of location, depth values and surface normals of pixels. Once homogeneous regions are determined, the top portion of the boundary of each region in the segmentation map is extracted and matched with realistic head-shoulder template curves. We evaluate our method both on a publicly available dataset, and on our new human detection dataset, which is composed of 500 depth images of humans in diverse poses acquired in varying indoor environments.

Introduction

Since the introduction of low-cost depth sensors such as Microsoft Kinect, research on scene analysis of indoor environments has gained momentum. Depth sensing provides rich information on surface geometry and unambiguous depth relationships between the objects and other structures present in the scene. Thus, depth information brings a powerful cue to segment objects of interest from the background.

Correct detection of humans is essential for tracking algorithms, especially for those that rely on detection at each frame. These algorithms search for the presence of humans in each frame separately, then relate detected human positions across video frames to provide tracking output. Apart from tracking algorithms, detecting unmoving humans in diverse, non-pedestrian poses from still depth images is especially important for robots performing domestic tasks or office service, for the care of sick, elderly, or disabled people, and for indoor search and rescue operations after a disaster.

Our main objective is to detect humans in indoor environments from depth images without assuming the following constraints:

- There is no constraint on the number of people present in the scene.
- The humans are not assumed to be moving. Hence, the detection algorithm does not employ approaches, such as background subtraction, that rely on motion information.
- There is no restraint on the environment as long as the humans are in the range of the depth sensor. There can be any amount of clutter in the scene.
- The only assumption on the pose of the people is that their head and shoulder regions are at least partially visible. Humans can be seen in pedestrian poses, i.e. standing and walking, or non-pedestrian poses, such as lying down, fallen, etc.
- As long as head and shoulder region is not completely occluded, the humans can be seen occluded by other objects or humans.
- The ground plane is not assumed to be visible. The ground plane may be cluttered with furniture and other objects, or the ground may not be in the field of view of the sensor. Besides, humans can be on sofas, beds, or armchairs.

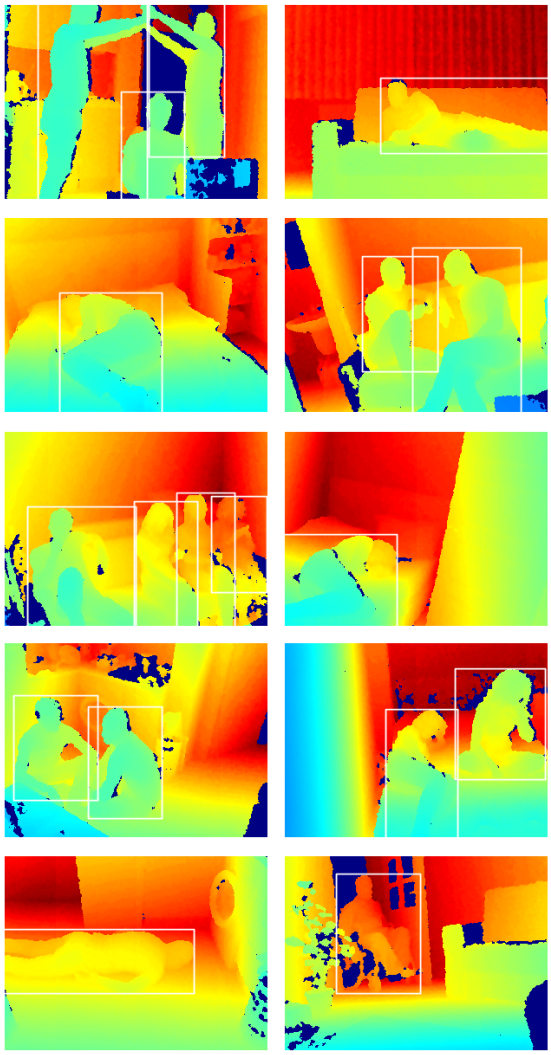

Figure 1. *Samples from our new human detection dataset. White rectangles indicate ground truth bounding boxes.*

The object-background or object-object discontinuity in a depth image can manifest itself either as depth discontinuity or surface normal discontinuity, or both. We propose a method that exploits the homogeneity of depth values and surface normals

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

over regions corresponding to distinct objects. To this effect, we apply K-means clustering to the depth pixels each represented as a vector of weighted location, depth value and surface normal. In addition to facilitate the search for candidate human occurences, the segmented regions provide additional information that may lead to a semantic interpretation of the environment; e.g. extraction of ground, walls, and furniture, as well as other objects of importance.

After segmentation, the boundary of a segmented region is examined to check whether its top portion resembles the shape of the head-shoulders part of the body, known as "omega-shape" in human detection literature. We locate potential neck positions at curvature minima, and then, we perform a full matching with realistic head-shoulder template curves to cover a variety of head-shoulder configurations.

In order to measure the effectiveness of our method, we constructed a dataset of 500 depth images containing humans in diverse poses (standing, sitting, sleeping, fallen, chatting, playing, etc.), acquired from various indoor environments (kitchens, living rooms, bedrooms, offices, classrooms, coffee shops, stores). The scenes contain a varying number of people with significant occlusion. We acquired the depth images with Microsoft Kinect Sensor for XBox 360. The human occurences are manually labeled with bounding boxes. See Figure 1, for sample images from our dataset.

Related work

In most of the previous work dealing with human detection from depth images, the main objective is tracking of moving humans. The people are assumed to be mostly in upright position, i.e. standing, walking, and sometimes sitting; poses which can be classified as pedestrian postures. Human detection approaches that assume such pedestrian detection/tracking framework can either rely on motion information to detect humans [15], [16] or use individual depth frames to localize human regions [3], [4], [6], [19]. Locating candidate regions using individual frames, and then associating candidates across video frames is a common approach in tracking, which greatly reduces both false positives and false negatives [5], [8], [14], [17], [20], [22].

Nizalowska et al [16] describe a tracking algorithm that uses motion information to determine candidate human silhouettes. The top of a silhouette is marked as a head region if the region conformed with a set of rules of "headness". Nghiem et al [15] apply background subtraction to detect moving objects, and then search for head locations restricted to these moving objects. They use the assumption that the head contour is nearly a full circle. They also use HOG descriptors to further classify human regions from non-human ones. However, these methods, which require motion information to detect humans are not applicable to situations where humans are not moving in a domestic or office environment, or after a disaster.

In some of the methods developed to detect human candidates from individual depth frames, it is assumed that the ground or the ceiling is visible [3], [14], [17], [22]. Bagautdinov et al. [3] searched for the presence of objects/humans on a ground plane using Bayesian inference. In [14], a method is proposed to track walking people from RGB-D images. The ground plane is detected and removed from the point cloud data, and the remaining points are clustered to determine candidate human regions. Likewise, Zhang et al. [22] used RANSAC to detect the floor and the ceiling in the point cloud to extract clusters from the remaining point cloud that might correspond to human regions. Then a HOG-based people detection method is applied on the clustered regions. This approach is based on the assumption that humans stand isolated in an upright position, and the environment is not cluttered. In our method, we avoid the assumption that a ground or ceiling plane is visible in the scene, since that does not always hold in domestic environments where the scene is cluttered with furniture and other home or office objects. Furthermore, humans are not always standing on the ground at home; for example, they are often seated or lying down on sofas (Figure 1).

In [6], the authors decomposed the point cloud into a set of layers at different heights, then clustered each layer separately. The clusters are classified into human or non-human via a random forest classifier. The histograms of local surface normals are used as features. The training and tests were performed on point clouds in which the subjects are standing, walking, sitting, and in some cases partially occluded. Jafari et al. [8] also assumed a pedestrian setting, where the point cloud is projected onto a plane to obtain a histogram of points and blob detection is performed on the projection to determine isolated objects. An upper body detector is applied to the local maxima of the depth values in a blob to localize human occurences.

In [20], Xia et al. developed a human detection method that aims at first detecting heads from the edge map of depth images. They use Chamfer distance to a template head contour to find candidate heads, then verify using a 3D head sphere model. To reduce the high false positive rate, association between successive video frames is used. Choi et al [4] employed a graph-based segmentation algorithm followed by a region merging step to determine candidate human regions. Then they compute Histogram of Depth (HOD) descriptors on the candidate regions and classify the regions using linear SVM. Choi et al. [5] developed an elaborate human tracking system that applies various cues and detectors to video data. Depth information can be integrated in this system when available. Their depth-based shape detector employs a binary head-and-shoulder template to evaluate the likelihood of human presence in target locations.

In some other work, metric distances obtained through the calibration of the depth sensor are used to eliminate regions that do not correspond to the geometric configuration of standing humans [18]. Spinello and Arras [18] developed Histograms of Oriented Depth (HOD) descriptor where gradients of depth image are encoded. The HOD descriptor is combined with Histogram of Gradients (HOG) of RGB image to be classified using SVM. They computed and classified the descriptors only on windows whose depth values conform with the proportions of standing humans.

The head-shoulder contour has a distinctive shape that is salient for most human poses and view angles. This distinctive shape is widely exploited for human detection, tracking, and head-shoulder contour estimation from 2D intensity/color images, and is referred to as "omega-like shape" [7], [9], [10], [12], [13], [21]. In most of these methods, a classifier is trained on features extracted from rectangular image patches that contain head and shoulders. In our work, instead of using a sliding window detector trained on features extracted from head-shoulder image patches, we directly extract the head and shoulder contour from segmented regions, since depth and surface normal discontinuities in depth

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

images provide a robust object-background separation. Mukherjee and Das [12], [13] have also extracted contours from candidate regions to search for the presence of an omega shape; however they determine candidate regions via adaptive background modeling in color videos. Although we work on static depth images rather than RGB images, and we perform a full segmentation over the image, our approach is similar to [13]. The main difference is that we extract the omega-shape candidate via determining neck positions at curvature minima, and then, we perform a full matching with realistic head-shoulder template curves to cover a variety of head-shoulder configurations.

Existing depth datasets constructed for human detection/tracking assume a pedestrian detection framework. The images are acquired mostly in non-domestic environments, such as university halls, offices, and laboratories. The Kinect Tracking Precision Dataset (KTP) constructed by Munaro et al. [14] consists of RGB-D sequences of people walking in a lab. The RGB-D People Dataset [11], [18] contains more than 3000 RGB-D frames acquired in a university hall from three static Kinect cameras. The people in the dataset are in mostly upright position, walking or standing. The UTKinect-HumanDetection Dataset [20] is a sequence of depth images taken by the Kinect sensor for XBOX 360. The sequence contains 98 depth images of two people walking in a lab environment. The Kinect Office Dataset, collected by Choi et al. [5] is composed of 17 videos, each 2 to 3 minutes long, acquired in an office from a static Kinect camera. People can be found in sitting or standing positions. The 18 video sequences of the Kinect mobile dataset [5] were obtained from a Kinect sensor mounted on a mobile robot that wandered in the offices, corridors, hallways, and cafeteria of an office building. The EPFL-LAB and EPFL-CORRIDOR datasets described in [3] are created to track multiple people in upright positions. In contrast to these datasets, our new dataset contains scenes from various home and office environments, as well as stores and coffee shops, with people being in both pedestrian and non-pedestrian poses.

Human detection from depth images

Our human detection algorithm is composed of three main steps: First we segment the depth scene using K-means clustering, and merge adjacent planar regions. Prior to segmentation of the depth image, the zero depth values are filled by the algorithm developed in [2]. The second step is extracting omega-like curves from the top portions of the boundaries of the segmented regions, and matching them with template head-shoulder curves. Finally, the candidate head-shoulder regions are examined to check whether they satisfy two geometrical constraints attributed to valid head and shoulder regions.

Segmentation of depth images

The depth scene is segmented into homogeneous regions in terms of proximity, depth and surface normals. To this end we use a scheme that is based on K-means clustering of position, depth, and surface normal data, followed by median filtering and connected components algorithm. We also employ a region merging method to extract large planar regions. The resulting segmentation map is suitable for further processing to infer other objects of interest and structures, i.e. the ground plane, walls, and furniture.

Let d be the depth value at pixel location (x, y), and $[N_x \ N_y \ N_z]$ be the unit surface normal at that location. We normalize the coordinates and depth values by dividing them with their corresponding maximum values in the depth image to obtain $\hat{x}$, $\hat{y}$, and $\hat{d}$. The components of the normal vectors remain in the range of $[-1, 1]$. For each pixel we form the weighted vector $\bar{p} = [w_{xy}\hat{x} \ w_{xy}\hat{y} \ w_d\hat{d} \ w_N N_x \ w_N N_y \ w_N N_z]$. The weighted vectors of all pixels are then clustered into K clusters via K-means algorithm. The weights, w_{xy}, w_d, and w_N determine the contribution of each component to the formation of clusters, controlling the proximity of the pixels, the similarity of depth values, and surface normal homogeneity in each cluster, respectively.

After K-means clustering, an index image is formed using the cluster membership of each pixel. Median filtering is applied to the index image to smooth the index image, i.e. to merge small regions formed due to surface normal noise, into their larger neighbors. Then, connected regions with the same cluster index are determined and each connected region is assigned a separate region identity.

Merging planar regions

The connected regions are further analyzed to extract large planar surfaces from the scene. Regions that are larger than a predetermined size are processed with RANSAC algorithm, and regions with a low proportion of outliers are classified as planar regions. Then, adjacent planar regions with similar surface parameters are merged into a single planar region.

Figure 2 gives some depth images from our dataset, along with the index images obtained with K-means clustering, and the segmentation maps after determining connected regions, and merging planar regions. The regions of the resulting segmentation map do not exactly correspond to semantic regions in the scene; however, there is considerable overlap as can be observed from Figure 2. This map is suitable for further processing to achieve a semantic interpretation of the environment.

Detection of head and shoulders

Each region R in the segmentation map that has an area larger than a threshold is treated as a candidate human region. The boundary of the region is extracted and then smoothed with a gaussian filter of standard deviation σ. The topmost point (x_t, y_t) of the smoothed boundary is marked as the top point of a potential head. The curvature of each point on the boundary contour is computed and the local minima of the curvature values are determined. If a local minimum to the left of the topmost point is below a certain threshold c_t, then this point is treated as a potential neck point (x_{nl}, y_{nl}). Suppose that the number of points between the topmost point (x_t, y_t) and the left neck point (x_{nl}, y_{nl}) is L (Figure 3). Then a curve is formed that is composed of $2L$ points to the right of (x_t, y_t) and $2L$ points to the left of (x_t, y_t), together with the topmost point itself. This new curve is saved as a candidate head and shoulders curve. If a local minimum to the right of the topmost point also falls below the threshold c_t, another candidate head and shoulders curve is created using this right point (x_{nr}, y_{nr}). If the smoothed boundary of the region does not have local minima below the threshold, then the region is discarded. Head and shoulder curves that are too short or too long are also discarded.

Let us call a candidate head and shoulder curve $C = \{x_i, y_i\}$, where $i = -2L, .., 2L$. The coordinates of the candidate curve are normalized such that the topmost point is at the origin, and they

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

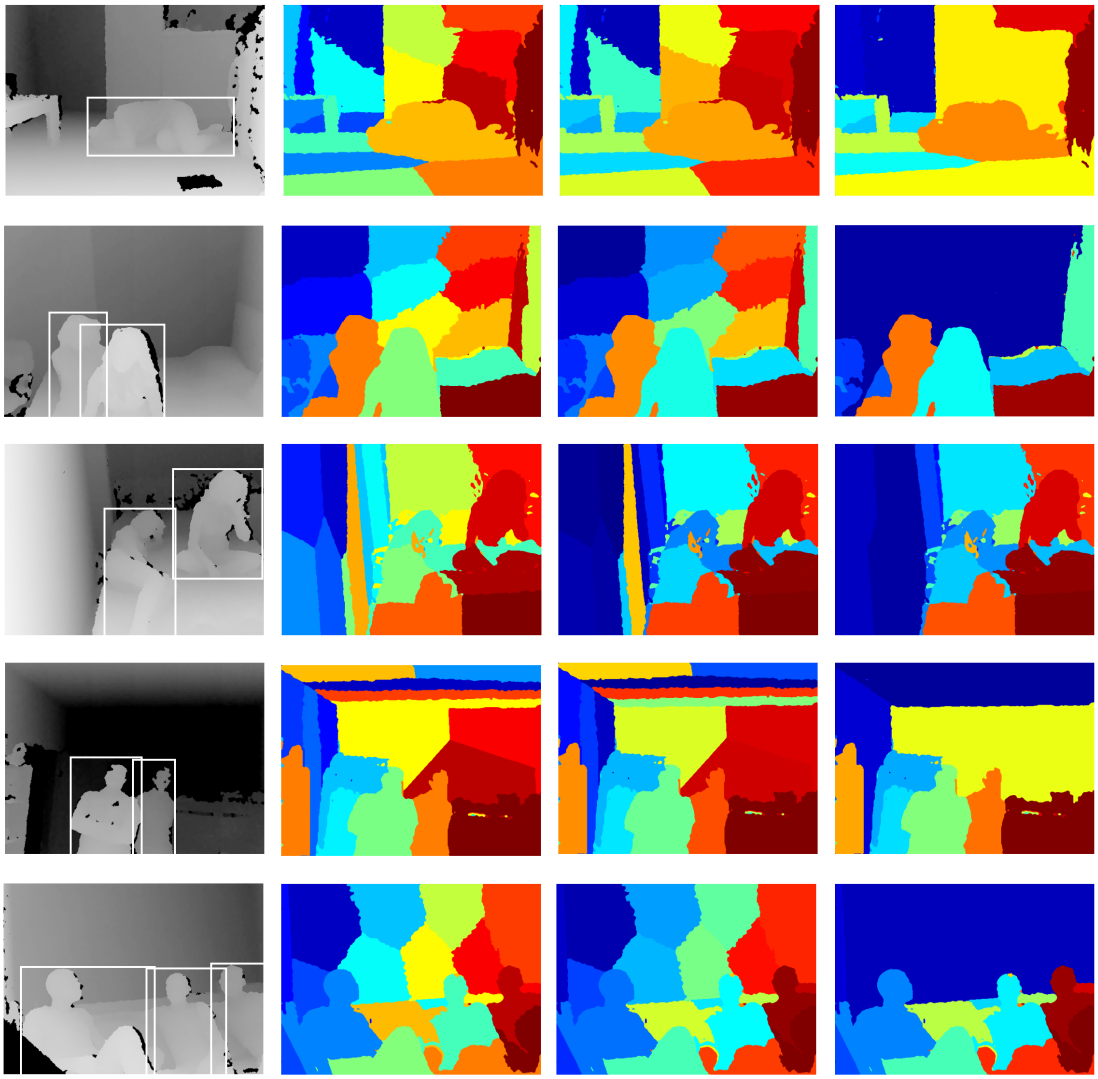

Figure 2. *First column: Depth images from our new human detection dataset. White rectangles indicate ground truth bounding boxes for humans. Second column: Index image obtained with K-means clustering. Third column: Regions after determining connected components. Fourth column: Segmentation map after merging planar regions.*

are scaled such that the average Euclidean distance between the topmost point and the two neck points becomes one. The curves are also rotated such that the bisector of the angle formed by the topmost point and the neck positions to the right and left becomes vertical. This step accounts for the slight rotations of the head and shoulder region. After this normalization, the normalized candidate curve $\bar{C}$ is matched with a set of head-shoulder templates via Hausdorff distance. Let $d(\bar{C}, T_m)$ be the Hausdorff distance between the candidate contour $\bar{C}$ and the mth template T_m. The candidate contour is eliminated if $D = min_m \left(d(\bar{C}, T_m)\right)$ exceeds a threshold h.

The head-shoulder templates are not hand-tailored but are obtained from real data. They are extracted in the same manner as described above from depth images acquired with Kinect sensor, and checked manually whether they truly are head and shoulder curves. These head and shoulder templates belong to humans who are not present in our dataset. Figure 4 shows some examples of such curves. The rich collection of templates enhances the ability of the system to capture a variety of head and shoulder configurations. The disadvantage is an increase in false detection rates since many non-human regions have top contours that are similar to these template curves.

Verification

In order to reduce the false detection rate, we include a simple verification step that checks whether the candidate head-shoulder regions satisfy two constraints: 1) The depth values just above the head should be larger than those within the head region. 2) The product of the perimeter and the median depth of the head and shoulder region should be within a certain range.

The first constraint is based on the observation that the pixels above a visible human head should have higher values of distance

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

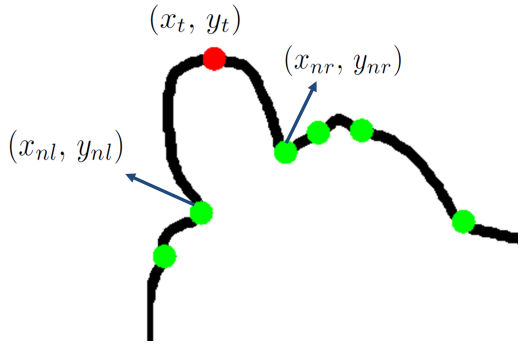

Figure 3. *The topmost point (in red), and the curvature local minima (in blue) on a boundary contour. The right (x_{nr}, y_{nr}) and left (x_{nl}, y_{nl}) neck points are also indicated.*

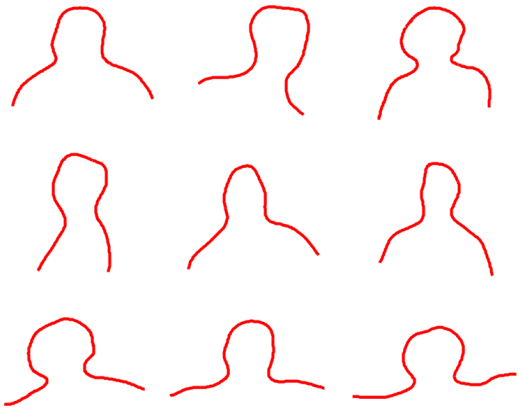

Figure 4. *Some examples of head-shoulder templates.*

from the camera, i.e. the head is assumed to be in front of the objects corresponding to the pixels just above the head contour. We define a rectangle centered at the topmost point of the head. The height of the rectangle is twice the average distance from the topmost point to the two neck points of the contour. The width is set to be one fourth of the height. This rectangle centered at the topmost point of the head is divided two equal parts as the upper part and the lower part. If the average depth values inside the upper part of the rectangle is higher than the average depth values of the lower part by a certain amount, then the head region is verified; otherwise it is discarded.

The second constraint limits the size of the head-shoulder region given its median depth value. We compute the products of the perimeters and the median depths of head-shoulder regions in a training dataset. In a test image, if the product is close to the average value by a constant multiple (q) of the standard deviation, then the head-shoulder region is verified. Let the median depth within the head and shoulder region be d_{HS}, the perimeter of the region be p_{HS}, μ_{dp} and σ_{dp} be the average and standard deviation of the depth perimeter product in the training set. The region is verified as a head-shoulder region if the condition

$$|d_{HS}p_{HS} - \mu_{dp}| < q\sigma_{dp} \tag{1}$$

is met.

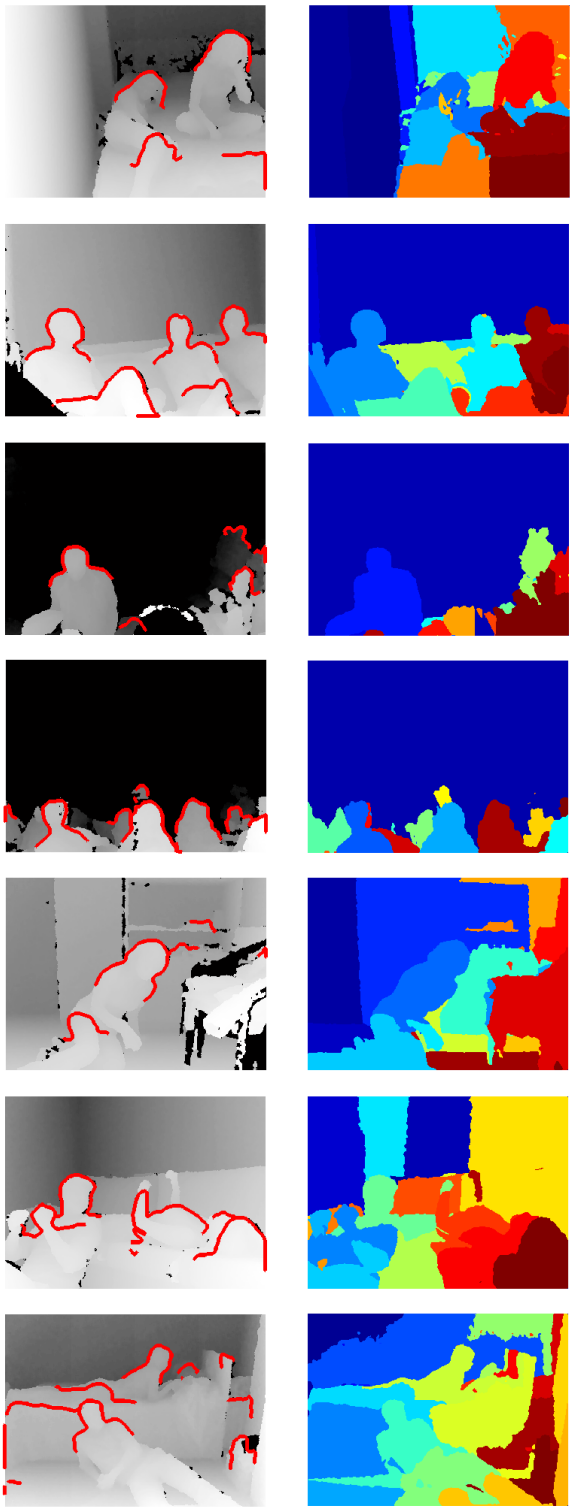

Figure 5. *Detection results from our dataset. On the left, curves in red indicate the detected head-shoulder curves. On the right, the segmentation maps are given.*

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

New Human Detection Dataset

We created a new dataset of 500 depth images of 1016 humans in total acquired by Kinect sensor for XBOX 360. In contrast to existing datasets of depth images [3], [5], [14], [18], [20], which are generally designed for tracking purposes, our dataset does not contain frames of video sequences. Instead, it consists of still depth images where motion information is unavailable for human detection. The images were collected in various indoor environments, such as living rooms, kitchens, bedrooms, offices, home offices, classrooms, corridors, stores, and coffee-shops. The humans in the images are diverse in identity and in poses. There is no restriction on the pose, humans can appear standing, walking, dancing, sitting, crouching, lying down, etc. A scene can contain a number of humans that are occluded by other humans or objects. The ground truth is obtained through manually marking the bounding box corresponding to each human in the scenes. The dataset, which we call ESOGU RGB-D Human Dataset, is publicly available on our web page along with the ground truth[1]. Figure 1 shows some example images from our dataset, with the ground truth bounding boxes marked as white rectangles.

In our experiments, we separated the dataset into a training and test set, composed of 200 and 300 images, respectively. The training set is used to set the parameters for the segmentation and head-shoulder curve extraction steps. In the training set, 354 humans are present in total, while in the test set 662 human occurrences are seen.

Experimental Results

We evaluated the performance of our human detection method on two datasets. The first is our new dataset where we used 200 images for observing the effects of segmentation related parameters, and the rest of the 300 for testing. The second dataset is the UTKinect-HumanDetection Dataset [20], which is a sequence of depth images taken by the Kinect sensor for XBOX 360. The sequence contains 98 depth images of two people taken in a lab environment.

Evaluation

We measure the performance of our algorithm with precision and recall, which are based on the number of True Positives (TP), False Negatives (FN), and False Positives (FP) returned by the algorithm. The precision and recall are defined as follows:

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Given a depth image, our human detection algorithm returns a number of head-shoulder contours. For the UTKinect-HumanDetection dataset, the ground truth is given as a point indicating the center of a head. If a detected contour encloses a ground truth point, it is marked as a true positive. For our new dataset, where the ground truth is given in terms of bounding boxes, if a detected contour is enclosed by a ground truth bounding box, and if the top of the contour is close to the upper edge of the bounding box by at least 10 pixels, the detected contour is counted as a true positive, otherwise it is a false positive. The undetected ground truth instances are counted as false negatives.

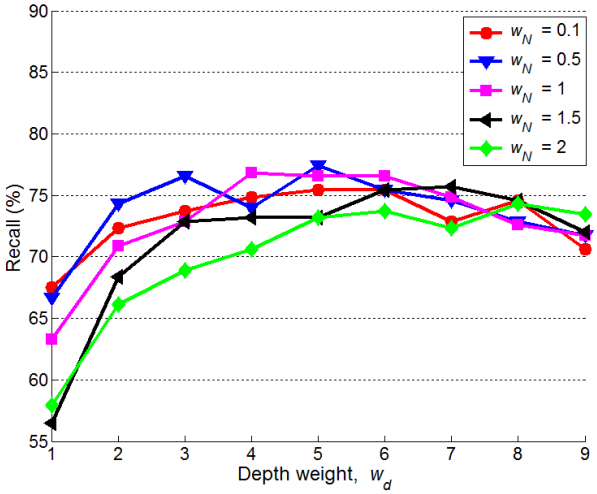

Figure 6. *Recall (%) yielded by various combinations of w_d and w_N.*

Results on the training images

We used our training dataset composed of 200 images with 354 human occurrences to observe the effects of the following parameters: 1) The weights for the position, depth, and normal components that are clustered via K-means algorithm; 2) The standard deviation of the Gaussian filter that is used to smooth region boundaries.

In our experiments, we kept w_{xy} as 1 and varied w_d and w_N to get the combination that gives the best recall. Figure 6 gives the results for various combinations of w_d and w_N. While obtaining these results, we set σ and h as 4 and 1, respectively. The number of head-shoulder template curves is 20. We eliminated the verification step via Equation 1 at this stage. As can be observed from the graphs in Figure 6, the performance is stable when w_d is between 4 and 6. In this range w_N should be set as less than 1. The contribution of depth values for accurate segmentation is higher than that of the surface normal components. The best recall value for the training set is 77.4 % with $w_d = 5$, and $w_N = 0.5$.

Table 1: Recall (R) and Precision (P) on the test dataset.

h	R (w/o)	P (w/o)	R ($q=6$)	P ($q=6$)
0.8	65.9	33.4	62.5	37.9
0.9	68.7	31.6	64.7	36.0
1	69.3	30.4	65.0	34.9
1.1	69.6	29.8	65.3	34.8
1.2	70.2	29.5	65.4	34.6

Figure 7 demonstrates the effect of σ on recall. The results are given for a set of best performing w_d and w_N combinations. The threshold on Hausdorff distance, h is set to 1. We can observe that $\sigma = 3$ and 4 are good choices for the standard deviation of the Gaussian filter used to smooth region boundaries.

Results on the test images

The test images are the remaining 300 images of our dataset. There are 662 human occurrences in these images. Setting the pa-

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

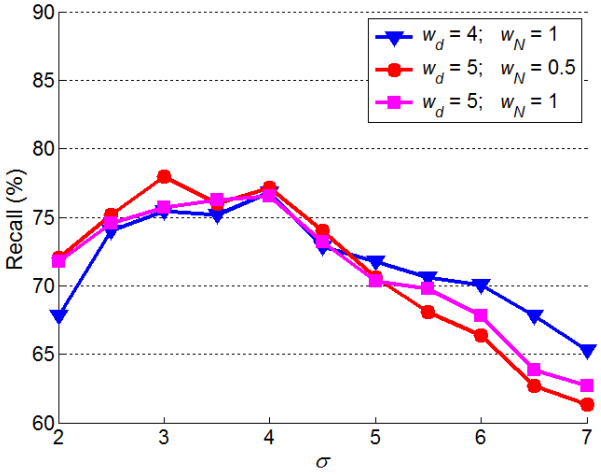

Figure 7. *Recall (%) with respect to σ.*

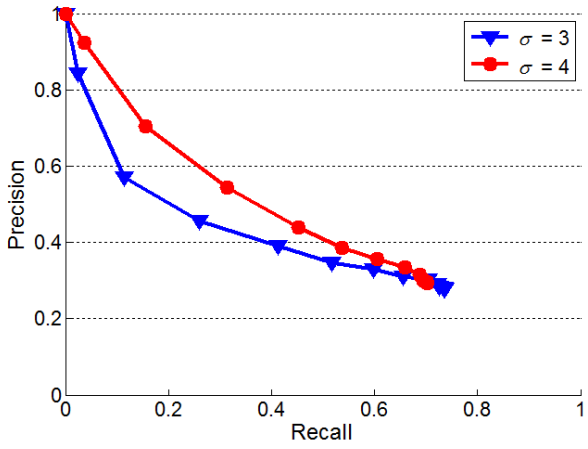

Figure 8. *Precision-Recall curves on test images for $\sigma = 3$ and $\sigma = 4$. The curves were obtained by varying h; the Hausdorff distance from the templates.*

rameters for segmentation as $w_d = 5$, and $w_N = 0.5$, we varied the threshold on Hausdorff distance from the templates (h) to obtain Precision-Recall curves. Figure 8 shows Precision-Recall curves for the best two choices for σ we had determined from the training set. We observe that, although with $\sigma = 3$ we achieve a higher recall (73.6 %) than with $\sigma = 4$ (70.2 %), the precision results are improved when σ is 4.

Up to this point we reported results we obtained without using the constraint in Equation 1. Figure 9 gives Precision-Recall curves when we impose this constraint on candidate head and shoulder curves that return a Hausdorff distance greater than 0.3. We give the Precision-Recall curves for the cases $q = 2$, 4, and 6, as well as the case when the constraint is not imposed. σ is set as 4 for these runs. Although the verification step reduces the false positives to a certain extent, it also has a large negative effect on recall.

In Table 1, precision and recall figures are reported with respect to a set of values of h and q. The algorithm is able to detect

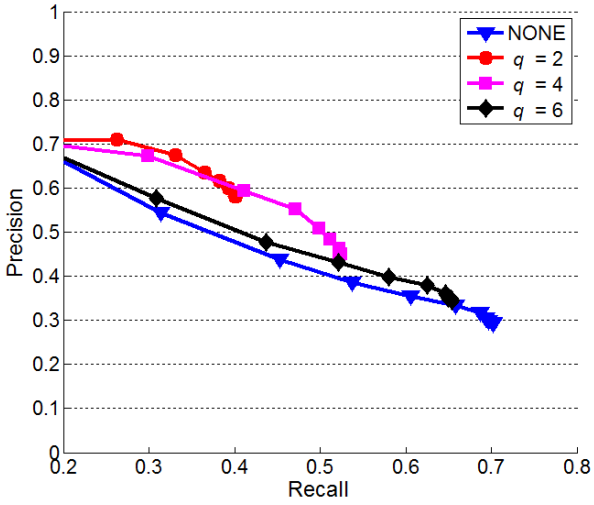

Figure 9. *Precision-Recall curves on test images for various q. The curves were obtained by varying h; the Hausdorff distance from the templates.*

70.2 % of the humans in the images with a high return of false positives (29.5 % precision and 3.7 false positives per image).

In Figure 5, we show some example depth images from our dataset. On the right column, the segmentation maps are given, and on the left the detected head-shoulder curves are marked in red on the depth images. The algorithm can extract head-shoulder regions with success; however it returns a large number of false positives. Some of these false positives are due to clutter, and imperfect segmentation. In some cases, the human body is further segmented, and body portions return boundaries that contain curves similar to head-shoulder templates.

Figure 10 shows cases where the algorithm failed to detect head-shoulder occurrences. In some cases, the segmentation algorithm separates the head region from the shoulders due to the large depth difference. In other cases, the head or shoulder regions merge with other structures, such as the wall or the bed. The algorithm also fails to locate the head and shoulder region, if it does not occur at the top of the region boundary.

Results on UTKinect HumanDetection Dataset

In the UTKinect-HumanDetection Dataset [20], which consists of 98 frames acquired in a lab environment, there are 176 occurrences of two persons. All the parameters, except for the threshold on Hausdorff distance, are as determined from our training set ($w_d = 5$, $w_N = 0.5$, $\sigma = 4$). We skipped verification via Equation 1 for the experiments on UTKinect-HumanDetection Dataset. The results are give in Table 2, along with the results reported in [20]. Notice that the performance figures in [20] are obtained through a tracking module that discards false positives via data association between video frames. Our algorithm missed 25 of the 176 positive instances. Most of the misses are due to occlusion of the head region.

Conclusion

We propose a human detection method that operates on still depth images. Our method is capable of detecting multiple hu-

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

Table 2: Results on UTKinect-HumanDetection Dataset.

	TP	FN	FP	R (%)	P (%)
h = 0.4	111	65	46	63.1	70.7
h = 0.9	151	25	333	85.8	31.2
XIA et al.[20]	169	7	0	96.0	100

mans in various poses and under significant occlusion, provided that their head and shoulders are visible. We present a new human dataset, which we constructed to test human detection algorithms operating on single frames. We tested our algorithm on this new dataset and a publicly available dataset. For both datasets, our algorithm is able to detect over 70% of the humans present in the scenes.

Acknowledgments

This work is supported by Eskisehir Osmangazi University Scientific Research Project (Project No:201415012(2013-253)).

References

[1] ESOGU RGB-D Human Dataset. http://mlcv.ogu.edu.tr/RGBDHuman.html.

[2] Smoothing kinect depth frames in real-time. http://www.codeproject.com/Articles/317974/KinectDepthSmoothing.

[3] T. M. Bagautdinov, F. Fleuret, and P. Fua. Probability occupancy maps for occluded depth images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2829–2837, 2015.

[4] B. Choi, Ç. Meriçli, J. Biswas, and M. M. Veloso. Fast human detection for indoor mobile robots using depth images. In *2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*, pages 1108–1113, 2013.

[5] W. Choi, C. Pantofaru, and S. Savarese. A general framework for tracking multiple people from a moving camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1577–1591, 2013.

[6] F. Hegger, N. Hochgeschwender, G. K. Kraetzschmar, and P. Plöger. People detection in 3d point clouds using local surface normals. In *RoboCup 2012: Robot Soccer World Cup XVI [papers from the 16th Annual RoboCup International Symposium, Mexico City, Mexico, June 18-24, 2012]*., pages 154–165, 2012.

[7] J. C. S. Jacques and S. R. Musse. Improved head-shoulder human contour estimation through clusters of learned shape models. In *28th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2015, Salvador, Bahia, Brazil, August 26-29, 2015*, pages 329–336, 2015.

[8] O. H. Jafari, D. Mitzel, and B. Leibe. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, May 31 - June 7, 2014*, pages 5636–5643, 2014.

[9] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. In *19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*, pages 1–4, 2008.

[10] M. Li, Z. Zhang, K. Huang, and T. Tan. Rapid and robust human detection and tracking based on omega-shape features. In *Proceedings of the International Conference on Image Processing, ICIP 2009, 7-10 November 2009, Cairo, Egypt*, pages 2545–2548, 2009.

[11] M. Luber, L. Spinello, and K. O. Arras. People tracking in RGB-D data with on-line boosted target models. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, 2011*, pages 3844–3849, 2011.

[12] S. Mukherjee and K. Das. A novel equation based classifier for detecting human in images. *CoRR*, abs/1307.5591, 2013.

[13] S. Mukherjee and K. Das. Omega model for human detection and counting for application in smart surveillance system. *CoRR*, abs/1303.0633, 2013.

[14] M. Munaro and E. Menegatti. Fast RGB-D people tracking for service robots. *Auton. Robots*, 37(3):227–242, Oct. 2014.

[15] A. Nghiem, E. Auvinet, and J. Meunier. Head detection using kinect

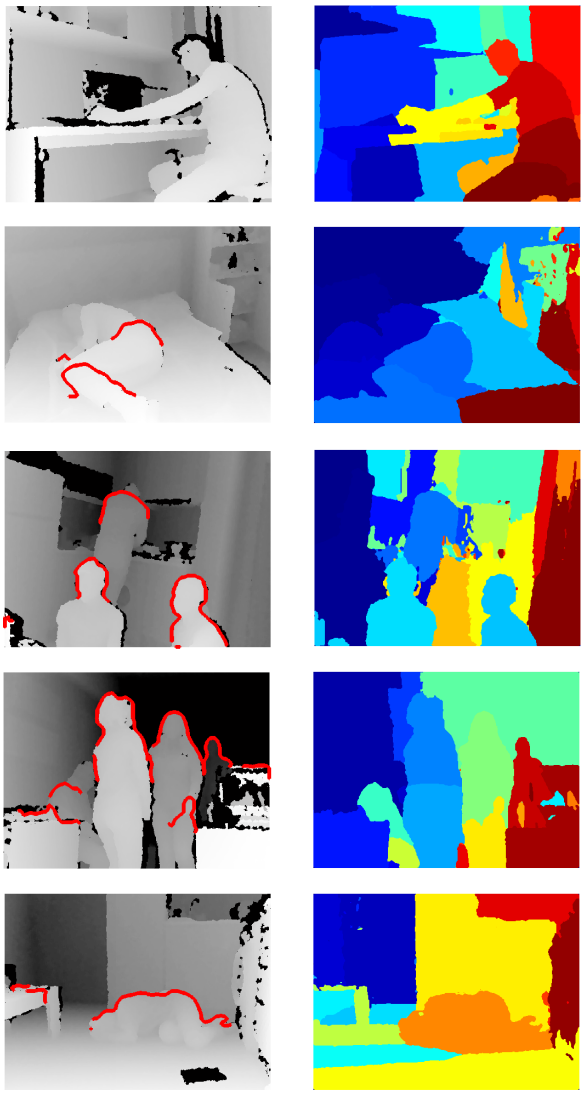

Figure 10. *Detection results from our dataset. On the left, curves in red indicate the detected head-shoulder curves. On the right, the segmentation maps are given. In these examples, missed human instances can be observed.*

DOI: 10.2352/ISSN.2470-1173.2016.21.3DIPM-046

camera and its application to fall detection. In *11th International Conference on Information Science, Signal Processing and their Applications, ISSPA 2012, Montreal, QC, Canada, July 2-5, 2012*, pages 164–169, 2012.

[16] K. Nizalowska, L. Burdka, and U. Markowska-Kaczmar. Indoor head detection and tracking on RGBD images. In *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, September 7-10, 2014.*, pages 679–686, 2014.

[17] J. Salas and C. Tomasi. People detection using color and depth images. In *Pattern Recognition - Third Mexican Conference, MCPR 2011, Cancun, Mexico, June 29 - July 2, 2011. Proceedings*, pages 127–135, 2011.

[18] L. Spinello and K. O. Arras. People detection in RGB-D data. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, 2011*, pages 3838–3843, 2011.

[19] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart. A layered approach to people detection in 3d range data. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, 2010.

[20] L. Xia, C. Chen, and J. K. Aggarwal. Human detection using depth information by kinect. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2011, Colorado Springs, CO, USA, 20-25 June, 2011*, pages 15–22, 2011.

[21] H. Xin, H. Ai, H. Chao, and D. Tretter. Human head-shoulder segmentation. In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, pages 227–232, 2011.

[22] H. Zhang, C. M. Reardon, and L. E. Parker. Real-time multiple human perception with color-depth cameras on a mobile robot. *IEEE T. Cybernetics*, 43(5):1429–1441, 2013.

Author Biography

Gulsum Nurdan Can recieved the B.Sc. degree in 2015 from the Department of Electrical-Electronics Engineering, Eskisehir Osmangazi University, Turkey.

Helin Dutagaci received the B.Sc., M.Sc., and Ph.D. degrees from Bogazici University, Istanbul, Turkey, in 1999, 2002, and 2009, respectively. She worked as a Guest Researcher at National Institute of Standards and Technology between 2008 and 2011. She is currently an assistant professor at the Department of Electrical-Electronics Engineering, Eskisehir Osmangazi University, Turkey. Her major field of study is signal processing. Her research interest includes computer vision, pattern recognition, 3-D object recognition, and biometrics.