Multiple Human Detection in Depth Images

Muhammad Hassan Khan*†, Kimiaki Shirahama*, Muhammad Shahid Farid†, Marcin Grzegorzek*
Research Group for Pattern Recognition, University of Siegen, Germany
{hassan.khan, kimiaki.shirahama, marcin.grzegorzek}@uni-siegen.de
†Punjab University College of Information Technology
University of the Punjab, Lahore, Pakistan
shahid@pucit.edu.pk

***Abstract*—Most human detection algorithms in depth images perform well in detecting and tracking the movements of a single human object. However, their performance is rather poor when the person is occluded by other objects or when there are multiple humans present in the scene. In this paper, we propose a novel human detection technique which analyzes the edges in depth image to detect multiple people. The proposed technique detects a human head through a fast template matching algorithm and verifies it through a 3D model fitting technique. The entire human body is extracted from the image by using a simple segmentation scheme comprising a few morphological operators. Our experimental results on three large human detection datasets and the comparison with the state-of-the-art method showed an excellent performance achieving a detection rate of 94.53% with a small false alarm of 0.82%.**

***Index Terms*—human detection, depth image, Template matching in the frequency domain**

I. Introduction

Human detection from images and videos is a challenging problem due to variations in body size, appearance, clothing, and background, etc. Although numerous RGB based human detection methods have been proposed and reported to achieve high accuracies [1]–[6], they suffer from various data capturing and processing artifacts e.g., color sensitivity, illumination variation, complex and cluttered backgrounds.

Recently, depth images are found to be very useful in human detection. The depth data has several advantages over the visual data as it provides 3D structural information of the scene which offers more details and important cues for human detection. Specifically, a depth image provides a layered structure of the scene where each pixel value represents the distance from the camera and can be used to extract foreground and background features more accurately. This layered structure enables us to accurately extract human regions even in the presence of a cluttered background. Although humans represented in depth images do not possess any color and texture information whatsoever, they occupy an integrated region in the image, robust to the change in color and illumination [7]. Moreover, the depth cameras may also work in low light and even in total darkness. Most depth base human detection algorithms encounter difficulties in perceiving the shape of humans when they are occluded with some other human or non-human objects which results in poor detection.

In this paper, we propose a simple and fast human detection algorithm in depth images obtained from Microsoft Kinect which can cope with the afore-mentioned problems. The proposed algorithm exploits the edges map of the depth image to detect humans. It searches a pre-defined head template in the depth edge map to find the human head. This template is able to detect the human head in all poses and from all angles. The matching between the edge image and the human head template is achieved by using the sum of squared differences (SSD) and cross correlation (CC) matching in an efficient way. We compute correlation in frequency domain instead of spatial domain which significantly reduces the computational cost. To verify that the detected region is a human head, it is fitted into a 3D head model which utilizes both the edge information and the relational depth change information from the original depth map. After verification, we apply a region growing algorithm on the head location to extract the entire human body. We also utilize morphological operators to improve the segmentation accuracy.

The human detection technique proposed in [8] finds the human from depth image by using a human head template and 2D chamfer matching. They exploit the distance transformation from the depth edge map to detect the human. Our approach is somewhat similar to [8], however, the proposed technique is different and novel in many respects which are summarized in the following:

- We exploit the relationship between SSD and CC matching algorithms which not only significantly improves the matching accuracy but also the implementation in frequency domain greatly improves the execution time of the technique.
- For human body segmentation, in addition to the segmentation filter, we propose to use morphological operators to improve its accuracy.
- Moreover, the proposed technique can be used to detect multiple people (two or more) in a scene while the method in [8] is limited to detect two persons only.
- The performance of the proposed algorithm is evaluated on three large datasets and the results show that the proposed algorithm is highly accurate.

The rest of the paper is organized as follows: Sect. II presents the related work. Sect. III describes the proposed multiple human detection and extraction technique. Experiments and results are reported in Sect. IV and Sect. V concludes the paper.

II. Related Work

A large number of existing algorithms exploit depth data in various ways to detect humans from images and videos. Shotton et al. [9] proposed a method to predict the 3D positions of body joints for human detection in a single depth image. They computed the depth comparison features of each pixel and classified into human body parts. This scheme was further extended in [10], [11] by aggregating votes from a regression forest and incorporating dependency relationships between body part locations, respectively. A deep learning based method is presented in [12] which utilized the appearance of different body parts and their spatial layout for human detection. In [13] a support vector machine has been trained to discriminate the human head and shoulders for detection purpose. Choi et al. [14] proposed a framework to track multiple peoples and to estimate camera's motion simultaneously using Markov Chain Monte Carlo method and particle filtering. For human detection, the framework exploits the weighted combination of seven different detectors to evaluate the observation likelihood. Combo-HOD [15] human detector exploits histogram of oriented gradient (HOG) and histogram of oriented depth (HOD). Like HOG in visual data, HOD encodes the direction of depth changes and relies on depth-informed scale-space search that leads to detection process. For classification, a linear SVM was trained and used to classify the respective descriptor. Munaro et al. [16] proposed a method for detecting and tracking of peoples in RGB-D data. Different clusters are calculated by applying voxel grid filtering on kinect depth stream and a height map is created for each cluster and human-head is detected in each individual map using local maxima.

Zhou et al. [17] proposed spatio-temporal matching (STM) between a 3D motion capture model and a set of point trajectories corresponding to high joint responses for the detection of humans in video data. The research presented in [18], [19] used a set of depth images and constructed a new reference image where each pixel is computed by taking the median value of several pixel values from the past images and found the typical human motion. These techniques may not be very effective if the human is static or some other objects are also moving at the same time.

The human detector proposed in [20] uses background subtraction to extract the objects. It then uses Haar-like filter based on a human model expressing the convex shape of shoulder-head-shoulder. Zhu et al. [21] proposed a human detection by tracking his head and torso in depth images. They proposed the circle and box fitting for the detection of head and torso respectively. In [22], human upper body template is proposed to slide over the depth image and detects a human region by using the Euclidean distance. All these techniques are computationally very expensive in finding the best match(es).

In contrast to the most existing human detection techniques which require prior training, the proposed algorithm does not require any such statistical learning, modeling or temporal information and therefore, it can be directly applied to any videos or still images. Moreover, the proposed algorithm exploits a fast template matching algorithm in frequency domain to efficiently detect and extract the human. To the best of our knowledge, it is the first time that frequency domain based matching in depth images is proposed for human-detection.

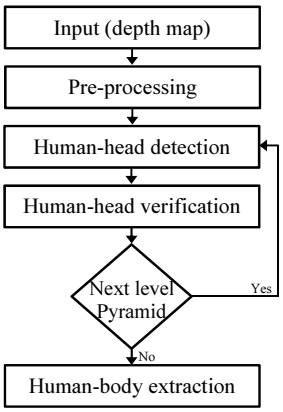


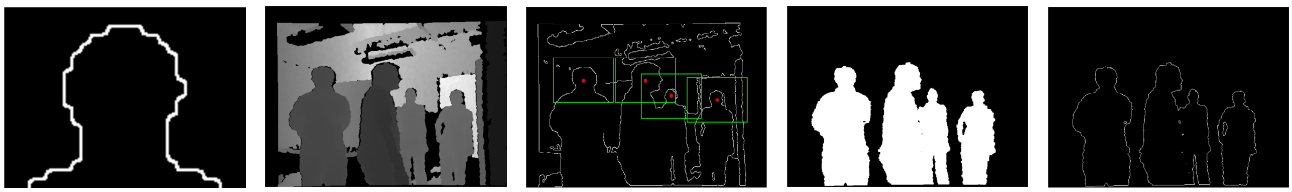
Fig. 1. Block diagram of the proposed multiple human detection technique.

III. Proposed Method

The proposed human detection technique works in three steps. In first step, the depth map is pre-processed to reduce the noise and recover the missing depth values. In second step, the human-head is detected and verified through head template matching and 3D model fitting technique respectively. In the final step, region growing algorithm is used on the detected and verified human-head location in original depth map to extract the human-body region with the help of a segmentation filter and few morphological operators. A block diagram showing the complete human detection process is presented in Fig. 1.

A. Pre-processing of Depth Maps

The 3D depth sensor in Kinect camera provides the depth information of a captured scene as a two dimensional (2D) array of pixels, known as depth map or depth image. Each pixel value of the depth map represents the distance (in millimeters) from the camera. For visualization and compression depth values are normalized in the range 0 to 255. The depth sensor in Kinect camera can only capture the distance information in the range 0.8 – 3.5 meters. If a pixel is out of this range, it is filled with the offset value 0 (zero). These pixels can be viewed as random black spots in the depth map and must be recovered to use the depth map in image processing techniques. We recover these missing pixels from the neighboring pixels by applying nearest neighbor interpolation algorithm. This converts the discrete data into continuous and recovers the depth value of missing pixels, however it may introduce some noise in the depth map. The noise can be removed by using order statistic filters. We used the median filter and tested it with various sizes to remove this noise and found the best results are obtained with a 5×5 size median filter. The reason



(a) (b) (c) (d) (e)

Fig. 2. Example of human detection using our proposed method. (a) the human-head template used in this paper, (b) a depth image containing four persons on various locations, (c) the human heads detected and verified using the proposed algorithm, (d) and (e) are the extracted silhouette and contour respectively.

of choosing the median filter is that it preserves the edges in the depth image, which are very important to the later stages of the proposed algorithm.

B. Human Head Detection and Verification

After pre-processing the depth map, the next step is to locate the position of the head in a depth map. The edges in the depth map are computed using the Canny edge detector [23]. The small edges do not contribute in head detection and therefore can be dropped to reduce the computational cost of the later stages of the proposed algorithm. The edges smaller than a predefined edge length τ are dropped. The edge map is used to find the head structures by comparing it with a head template. The predefined human head template used in this paper is shown in Fig. 2a.

For human head template matching in the edge image, the sum of squared differences (SSD) and cross correlation (CC) based matching algorithms are used [24]. Let f be the edge image and t be the human head template of size $M \times N$. In t, we express a pixel as $t(i,j)$ where i and j represent the x and y positions of the pixel, respectively. In addition, let us assume that we are now matching t with the rectangular region in the edge image of depth map f where a pixel in f is represented as $f(x,y)$. The SSD value at pixel (x,y) is computed as:

$$d(x,y) = \sum_{i=1}^{M}\sum_{j=1}^{N}\Big(f(x+i,y+j) - t(i,j)\Big)^2, \qquad (1)$$

SSD can also be viewed as squared Euclidean distance between the image patch of f and template t. Expanding Eq. 1 yields:

$$\begin{aligned} d(x,y) = &\sum_{i=1}^{M}\sum_{j=1}^{N} f^2(x+i,y+j) + \sum_{i=1}^{M}\sum_{j=1}^{N} t^2(i,j) \\ &- 2\sum_{i=1}^{M}\sum_{j=1}^{N} f(x+i,y+j)t(i,j) \end{aligned} \qquad (2)$$

In the above equation, the first part is the sum of squared values in the edge image, the second term belongs to the template image and the third term is twice the value of the correlation between the image patch and the template. Note that the term $\sum_{i=1}^{M}\sum_{j=1}^{N} t^2(i,j)$ is constant. Assuming that the term $\sum_{i=1}^{M}\sum_{j=1}^{N} f^2(x+i,y+j)$ (i.e., the local image energy) is approximately constant, the remaining term (i.e., cross correlation) is:

$$CC(x,y) = \sum_{i=1}^{M}\sum_{j=1}^{N} f(x+i,y+j)t(i,j) \qquad (3)$$

The template t is traversed over the entire edge image f from top-left to bottom-right and the correlation is computed at each pixel. The image patches with high correlation represent the possible matches (i.e., the possible existence of human-head). However, one should be careful while using a correlation to measure the similarity in intensity images because the correlation is high on the locations where the image intensity is high, even though they are not very similar. To overcome this problem, the normalized cross correlation can be used but this is not a problem in our case, as the proposed method exploiting only binary images which contain the edge information (i.e., $f(x,y)$ & $t(i,j) \in \{0,1\}$).

Cross correlation can be computed by taking the inner product of the template and image patch, and it is computationally expensive [25]. For a search area of the size $M \times M$ and a template of the size $N \times N$, Eq. 3 requires approximately $N^2(M-N+1)^2$ additions and $N^2(M-N+1)^2$ multiplications [26]. In order to reduce the computational overhead, we compute the correlation in frequency domain instead of spatial domain. We use the Fourier frequency transform and exploit the correlation theorem which states that multiplying the Fourier transform of one function by the complex conjugate of the Fourier transform of the other gives the Fourier transform of their correlation.

$$CC(u,v) = \mathcal{F}^{-1}\Big(\mathcal{F}^*(f(x,y))\mathcal{F}(t(i,j))\Big) \qquad (4)$$

where $\mathcal{F}$ represents the Fourier transform, $\mathcal{F}^*$ represents the complex conjugate and $\mathcal{F}^{-1}$ is the inverse Fourier transform. The computational complexity of the correlation computation in frequency domain (Eq. 4) with a search area of the size $M \times M$ and a template of the size $N \times N$ is $12M^2 \log_2 M$ real multiplications and $18M^2 \log_2 M$ real additions [26]. Correlation computation in frequency domain is 2.5 to 12 times faster than in spatial domain [25].

The size of the head varies from person to person and depends on the distance of the person from the camera as well. A person appearing close to the camera is characterized by a large region of his head compared to the one at far distance. To handle such cases, we apply the proposed matching algorithm

(a) (b) (c) (d)

Fig. 3. Human body extraction: (a) a sample depth image, (b) intermediate results of proposed human detection algorithm on (a); the region growing algorithm also extracted few other objects connected with the detected human body, (c) results after applying the gradient filter on depth image before applying the region growing algorithm, (d) final results of the proposed algorithm after applying the morphological operator and estimating the missing pixels through neighborhood interpolation.

algorithm in a multi-resolution fashion which makes the proposed algorithm robust to scale change. We use the sampling rate of $\frac{1}{4}$ for pyramids construction. Although, the number of pyramids levels actually depends on a captured scene, we have restricted up to level 4 to limit the computation cost with minimum effect on the performance.

As a result of template matching, a number of locations may be identified as possible head regions. However, all the detected regions may not contain human heads. Therefore, the detected regions are verified for human head through a 3D model fitting technique. A 3D hemisphere model is used for the verification of the detected regions. We use the 3D model fitting technique proposed in [8]. Fig. 2c shows the results of human head detection in Fig. 2b; the detected heads are shown in green rectangles and the red dots represents the matching locations.

C. Human Body Extraction

The complete human body is extracted from the depth map by using the detected and verified human head region on it. Since the depth of the detected human head is known, we can extract the whole human body by selecting the same depth pixels from the original depth map. However, this may also select few other non-human pixels having the same depth values (false positives). This can be handled by applying the region growing algorithm on the detected human head location instead of selecting the same depth pixels directly. The region growing algorithm is a pixel based segmentation method that compares a pixel to its neighboring pixels and grows itself by including them if their difference is less than a threshold Υ. The region growing algorithm may also generate some false positives if the human body is either occluded or connected by other objects having similar depth values. For example, if a person is standing on the floor or touching another object like a table, then the depth of the human body and the other connected objects would be the same. This will result in the extraction of the connected objects along with the human body as shown in Fig. 3b. This problem is solved by applying a gradient filter on the original depth map before using the region growing algorithm. The gradient filter helps us to extract the boundary between the human body and other connected regions. Applying such a filter solves the problem but it also produces some noise on the extracted region as shown in Fig. 3c. To remove this noise, we apply morphological operator 'dilation' to bridge the disconnected pixels. The remaining missing pixels are estimated though neighbor interpolation. Fig. 3d shows the results of the human extraction from the depth image Fig. 3a.

IV. Experiments and Results

The proposed human detection algorithm is tested on three datasets: Our collected dataset, Xia et al. dataset [8] and the Cornell activity dataset (CAD-60) [7]. All experiments presented in this section were carried out on a corei5 notebook computer with 2.60 GHz CPU and 8GB RAM. We briefly describe each test database in the following.

A. Test Databases for Performance Evaluation

1) Our Collected Dataset: We collect a depth image database captured by using the Microsoft Kinect for evaluation of multiple human detection algorithms. The dataset consists of 1700 depth images containing 4378 human objects in total. The size of each depth image is 640×480. The dataset comprises depth images with single and multiple humans (up to 5) and particularly aims at evaluating the performance of multiple human detection algorithms. It also contains depth images in which the humans are occluded by other objects or humans, or interacting with other surrounding objects e.g., chair, table, computer and shelf. We also captured the periodic movements of people including walking, sitting, and standing.

2) Xia Dataset: This dataset consists of indoor images captured with Microsoft Kinect for XBOX 360. The dataset has two parts: the first part comprises 98 depth images containing 175 detectable human objects. Each image contains at most two humans with a variety of poses and interactions with other humans or surrounding objects. The second part of the dataset consists of 100 depth frames with 266 detectable humans and each frame contains at most 4 persons.

3) Cornell Activity Dataset: The third dataset we used to evaluate the performance of the proposed human detection algorithm is the well-known Cornell Activity Dataset (CAD-60)[1] [7]. The dataset consists of 60 RGB-D videos ($80,312$ frames) of four subjects (two males and two females) performing twelve unique activities: rinsing mouth, brushing teeth, wearing contact lenses, talking on the phone, drinking water,

[1] http://pr.cs.cornell.edu/humanactivities/data.php

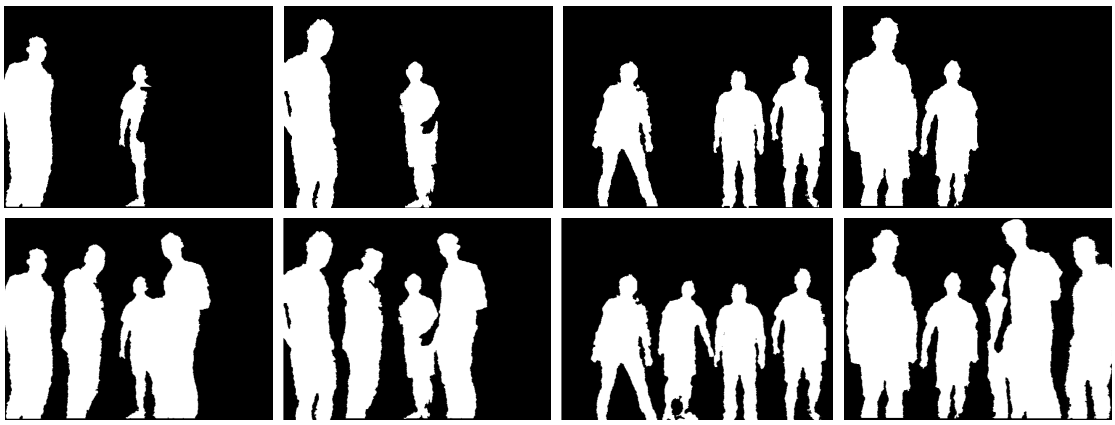


Fig. 4. Human detection results of Xia's algorithm [8] (top row) and proposed algorithm (bottom row) on four sample depth images from our collected dataset.

opening pill container, cooking (chopping), cooking (stirring), talking on the couch, relaxing on the couch, writing on a white board, and working on the computer. Although the dataset contains color images with corresponding depth maps and skeleton data, we used only the depth information in evaluation of proposed algorithm.

The CAD-60 dataset provides the tracked skeleton for each frame, however the ground truth of the segmented humans is missing which is required for objective evaluation of human detection and segmentation algorithms. After evaluating the proposed algorithm on CAD-60, the extracted silhouette and contour of objects in each frame were carefully analyzed and manually corrected, and we prepared the ground truth of CAD-60 dataset. The ground truth data will facilitate the research community in statistical evaluation of their research on CAD-60 dataset.

B. Performance Evaluation

The performance of the proposed human detection algorithm is evaluated on three large datasets described in the previous section. The results in terms of detection rate are reported in Tab. I. We also measure the performance of the proposed method in terms of precision, recall and accuracy. We have only a few FP (false positives) and FN (false negative) instances where the human head is either not visible or it is largely occluded by another human or non-human object in the scene. The results are presented in Tab. I which shows that the proposed algorithm is highly accurate. In particular, it achieves the detection rate of 94.53% with a small false alarm of 0.82%, precision 99.11%, recall 94.01% and accuracy 93.22% over all the four test databases.

We also compared the performance of the proposed algorithm with the state-of-the-art Xia et al. [8] human detection algorithm on our collected dataset. The results are reported in Tab. II which reveals the superior performance of the proposed algorithm. Our algorithm outperforms the Xia's algorithm in all three performance parameters by significant margins. From experiments, we note that the Xia's approach performs pretty well when the humans are isolated in the frames, however its performance is very poor when the human is occluded by the surrounding objects. Our algorithm, on the other hand, performs equally well in both scenarios. Fig. 4 shows the detection results of the proposed algorithm and Xia's algorithm on four sample depth images from our collected dataset. It can be noted that in each depth image the Xia's algorithm failed to detect all the peoples, whereas the proposed approach effectively detected and extracted all the peoples present in the depth images.

In order to evaluate the computational gain we achieve due to the frequency based implementation of the proposed algorithm over its spatial domain based implementation, we execute both implementations on our collected dataset and compute the execution time. The results are listed in Tab. III, which show that the frequency based implementation is approximately 47 times faster than the spatial domain implementation of the proposed algorithm. Tab. III also reports the average execution time of Xia's algorithm which is 1.5 times slower than the proposed algorithm.

TABLE I
ACCURACY OF PROPOSED METHOD

Dataset	Detection Rate	Precision	Recall	Accuracy
Our collected	98.58	99.61	98.58	98.20
Xia (Part-I)	98.29	99.42	98.29	97.73
Xia (Part-II)	94.36	98.05	94.36	92.62
CAD-60	94.30	99.09	93.75	92.95
Over all	94.53	99.11	94.01	93.22

TABLE II
PERFORMANCE COMPARISON OF PROPOSED ALGORITHM WITH OTHER TECHNIQUES.

Detection Method	Precision	Recall	Accuracy
Proposed	99.61	98.58	98.20
Xia [8]	90.05	85.01	77.71

TABLE III
COMPUTATIONAL TIME COMPARISON ON OUR COLLECTED DATABASE. TIME IS AVERAGE EXECUTION TIME (SECONDS) PER FRAME.

Algorithm	Time
Proposed (Frequency Domain)	2.47
Proposed (Spatial Domain)	116.09
Xia [8]	3.83

V. CONCLUSION AND FUTURE RESEARCH

In this paper, a template based human detection approach in depth images is proposed. The proposed technique works in three steps. In first step, the depth maps are pre-processed to recover the missing depth values. In second step, the edges from depth data are computed and used to detect the human-head through a frequency domain based template matching. The detected human-head region is verified through a 3D model fitting technique. Finally, the entire human body is extracted from the depth images using the region growing algorithm and the proposed segmentation technique. The results of experimental evaluation on three datasets and performance comparison with the state-of-the-art method show that the proposed method can effectively detect and extract multiple persons with various appearances from the depth images. Moreover, the proposed algorithm does not require any prior training, and therefore it is applicable to any depth data without any modification. In future, we plan to develop a GPU based implementation of the proposed algorithm by including the detection of few other human-body parts if human-head is occluded and extend it to activity detection using the extracted silhouette or contours of the peoples.

ACKNOWLEDGMENT

This research was partially supported by the University of the Punjab, Lahore and the European Commission (Horizon 2020) within the project "My-AHA: My Active and Healthy Ageing" (Grant Number: 689592).

REFERENCES

[1] C. Thurau, *Proc. 2nd Workshop Human Motion – Understanding, Modeling, Capture and Animation*. Springer Berlin Heidelberg, 2007, ch. Behavior Histograms for Action Recognition and Human Detection, pp. 299–312.

[2] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul 1997.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, 1999, pp. 1150–1157 vol.2.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct 2005, pp. 1395–1402 Vol. 2.

[5] D. Toth and T. Aach, "Detection and recognition of moving objects using statistical motion detection and fourier descriptors," in *Proc. Int. Conf. Image Anal. Process. (ICIAP)*, Sept 2003, pp. 430–435.

[6] H. Sidenbladh, "Detecting human motion with support vector machines," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, vol. 2, Aug 2004, pp. 188–191 Vol.2.

[7] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proc. AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011.

[8] L. Xia, C. C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, June 2011, pp. 15–22.

[9] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, Jan. 2013.

[10] M. Sun, P. Kohli, and J. Shotton, "Conditional regression forests for human pose estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2012, pp. 3394–3401.

[11] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 415–422.

[12] M. Jiu, C. Wolf, G. Taylor, and A. Baskurt, "Human body part estimation from depth images via spatially-constrained deep learning," *Pattern Recognit. Lett.*, vol. 50, pp. 122 – 129, 2014, depth Image Analysis.

[13] M. Rauter, "Reliable human detection and tracking in top-view depth images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, June 2013, pp. 529–534.

[14] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1577–1591, July 2013.

[15] L. Spinello and K. O. Arras, "People detection in rgb-d data," in *Proc. Int. Conf. Intell. Robot. Syst. (IROS)*, Sept 2011, pp. 3838–3843.

[16] M. Munaro, F. Basso, and E. Menegatti, "Tracking people within groups with rgb-d data," in *Proc. Int. Conf. Intell. Robot. Syst. (IROS)*, Oct 2012, pp. 2101–2107.

[17] F. Zhou and F. De la Torre, *13th European Conf. Computer Vision (ECCV)*. Springer Int. Publishing, 2014, ch. Spatio-temporal Matching for Human Detection in Video, pp. 62–77.

[18] M. Kepski and B. Kwolek, "Unobtrusive fall detection at home using kinect sensor," in *Proc. 15th Int. Conf. Computer Analysis of Images and Patterns (CAIP)*. Springer Berlin Heidelberg, 2013, pp. 457–464.

[19] M. Kepski and B. Kwolek1, "Fall detection using ceiling-mounted 3d depth camera," in *Proc. Int. Conf. Computer Vision Theory and Applications (VISAPP)*, vol. 2, Jan 2014, pp. 640–647.

[20] S. Ikemura and H. Fujiyoshi, "Human detection by haar-like filtering using depth information," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Nov 2012, pp. 813–816.

[21] Y. Zhu and K. Fujimura, "Constrained optimization for human pose estimation from depth sequences," in *Asian Conf. on Computer Vision (ACCV)*, 2007, pp. 408–418.

[22] 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 *Proc. Int. Conf. Robotics and Automation (ICRA)*, 2014, pp. 5636–5643.

[23] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov 1986.

[24] J. Martin and J. L. Crowley, "Experimental comparison of correlation techniques," in *Int. Conf. Intelligent Autonmous Systems*, 1995, pp. 86–93.

[25] D. Lyon, "The discrete fourier transform, part 6: Cross-correlation." *Journal of object technology*, vol. 9, no. 2, pp. 17–22, 2010.

[26] J. Lewis, "Fast normalized cross-correlation," in *Proc. Vision interface*, vol. 10, no. 1, 1995, pp. 120–123.