Fusion of Thermal and Depth Image to Improve Human Segmentation for a Mobile Robot

Fusion of Thermal and Depth Image to Improve Human Segmentation for a Mobile Robot

H S Hadi[1,2], M Rosbi[3], and U U Sheikh[3]

[1]Department of Electrical Engineering, Politeknik Ibrahim Sultan, Johor Bahru, Johor, Malaysia
[2]Member of Malaysia Board of Technologies, Professional Technologies
[3]School of Electrical Engineering Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

saipolhadi@pis.edu.my

Abstract. In machine vision, surveillance systems are a kind of security that concentrates on the safety of the human and property. One of the main tasks of a surveillance system is the detection of humans. This paper presents a system of human detection and the development of a technique of human segmentation using a combination of information thermal and depth in a real indoor setting from a mobile robot. A novel fusion of thermal-depth information (FTDI) is introduced to enhance the efficiency of the segmentation process and expedite processing. In experimental studies, evaluation of the performance for the proposed system is carried out using Ground Truth (GT), in which the proposed system yield is compared to GT. The proposed system performs well with an approximate accuracy of over 90% for all data sets as illustrated in the quantitative results and even outperformed state-of-the-art algorithms. This paper presents the novelty of the work, in which the detection method can improve the classification of persons and their occlusion. The advantages, such as being computationally inexpensive and performs well even under severe occlusion and poor illumination, show that this proposed system is robust.

1. Introduction

A human can locate other persons under varying circumstances, regardless of colour or type of clothing, pose, appearances, partial occlusions, lighting or background disturbances, as mentioned by [1]. However, the analysis of colour and point of view of a person is not relevant to deciding that an object is a person, where the analysis of colour and point of view sometimes confuse between the person and other objects. Similarly, in surveillance systems, the above issues are most difficult and computationally intensive in human detection. In fact, it becomes more complex when these systems are using computationally intensive in the segmentation process.

Most researchers defined Region of Interest Generation (ROI) as the initial contour or rough contour of the object of interest, whereas others defined ROI as a rectangular area containing both the background information and object of interest. As mentioned by [2], ROI generation is assumed to contain persons if there are good correspondings with head-shoulders like a binary pattern. In [3], the generation of ROI is the first stage in image processing, and it is an essential step because it provides specific information for the following process. Fortunately, this stage can be supported by a particular hardware configuration.

The first stage of ROI generation is segmentation. Five general approaches to segmentation are Region-Based [4], [5], [6] and [7], Edge Based [8], Threshold [9], Feature-Based Clustering [10], and Model-Based [11]. Each approach has several techniques, and often researchers combine several approaches to improve the accuracy of segmentation what is already mentioned by approaches like [12] and [13]. The object segmentation stage consists of three consecutive phases. In the first phase, the image

is scanned to obtain possible candidates, and then the candidates will be filtered to remove any excess regions in the second phase. Finally, the image is clustered and labelled on the base of the candidate feature measurement, and each cluster is stored as a single object.

2. Previous Works

In the field of surveillance systems, detecting human existence is the main requirement. However, detecting humans is not as easy as it sounds because the human has a multifaceted appearance. In [14] used the "split-merge" approach to detect occluded persons. Firstly, the segmentation via over-segmented techniques is used for the foreground region of the depth data. Then the sub-regions produced from the above techniques are clustered into segments that are featured as 3D data. The depth-plane slicing is not to be implemented as the depth range of two persons in occlusion could overlap. Therefore, Wang's proposed a similarity-based metric of boundary for clustered sub-regions into segments. Unfortunately, the "split-merge" approach results in over-segmentation and sub-regions clustering techniques increase the computational cost, as shown in Figure 1. Furthermore, every region is segmented and clustered individually.

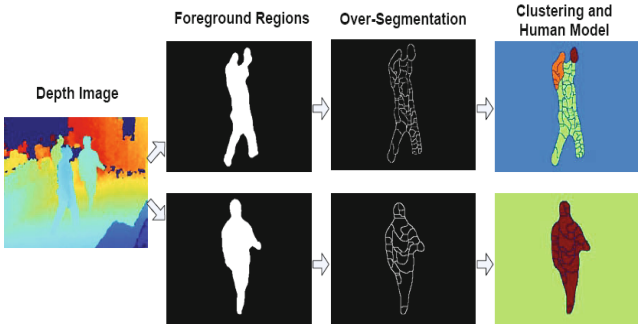

Figure 1. The proposed system for human detection by [13].

In [15] presented scientific achievements in mobile platforms for human detection and tracking using thermal vision. Kristoffersen's proposed fuzzy segmentation and genetic algorithm optimized threshold to solve the reflection of two or more persons crossing close to one another. Unfortunately, this system is inaccurate at detecting persons which are occluded with one another in a crowded scene and fail if the occlusion is too severe.

Work by [16] is one of the best detection approaches. They determine ROI in an image using a super-pixel method with the Edge-Boxes algorithm and then use Fast Region-based CNN (Fast R-CNN) model to extract features. Finally, the position of the pedestrian is detected. In [17] proposed a segmentation method to separate possible persons in the foreground using semantic segmentation network and network fusion architecture to integrate the pixel. This method is a reinforcement to the pedestrian detector. Unfortunately, the system has a significant loss in speed due to the complex network structure and large input size.

Fusion is a process of fusing the relevant modalities from several sources into a single image, where the results will be complete and more informative than any of the single image ([18], [19], [20], [21], [22]). These techniques can enhance the quality and increase the reliability of these data. Moreover, fusion techniques that can operate in real-time are more practical with the advances in the field of sensors, image processing algorithms, and high-performance computing technology. In [22] mentioned that most approaches could benefit from a fusion of modalities because it may reduce the miss detection that cannot be addressed by each individual modality.

In [23] presented a fast and accurate human detection method in complex environments from mobile robots using incorporated data of RGB and depth. However, this system does not address the problem of moderate to strong occlusion. In [24] presented a robust, accurate and fast method for human detection in the combination of RGB-Depth (RGBD) data that obtain extraordinary results compared to the state-of-the-art in human and pedestrian detection. Vaquero uses Shared Random Ferns for extracting features from RGB (HOG) and Depth (Histogram of Oriented Depth (HOD)) images. The system achieved an 89% detection rate including in complex environment situations.

However, work by [15] presented a pedestrian counting system with occlusion handling using stereo thermal cameras. This system was tested in a real-life scenario with heavy occlusion and moderate densities. The system performance is compared to the ground truth by manual annotation, and the system reached 95.4% and 99.1% counting rates for the two sequences. Work by [25] reported that there is not a single sensor that can reliably and effectively serve on mobile robots in a real-life scenario. This is because, once the robot moves, the background and the foreground become clutter. Pourmehr's use several modalities to counter the challenging real-life tasks like a microphone array (to detect the direction of sound sources), a laser range finder (to detect legs), and an RGB camera (to detect human torsos).

3. Methodology

First of all, this section was explained during the proposed system's procedure, as shown in Figure 2. The pre-processing stage is to deal with the data derived from the thermal and Kinect sensors. The raw data obtained from this sensor are manipulated to make the data suitable for the following stages. Region of Interest (ROI) Generation stage is the rough contour or initial contour of the object of interest in the thermal and depth images.

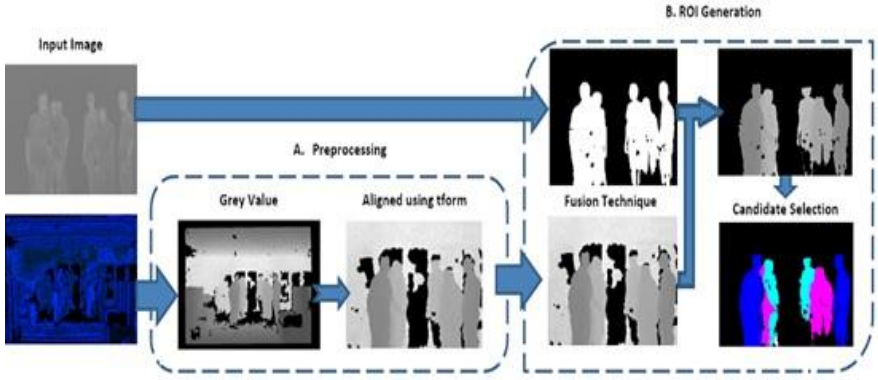

Figure 2. The procedure of the proposed system.

3.1. Pre-processing Stage

The pre-processing module will be conducted normalization and alignment of the depth data. The format of the recorded video for depth data (named *dpt*) is 24-bit. The real format of depth data is just an array of short, i.e. 16-bit, so it is necessary to convert the depth data into the real format (named *Idpt*).

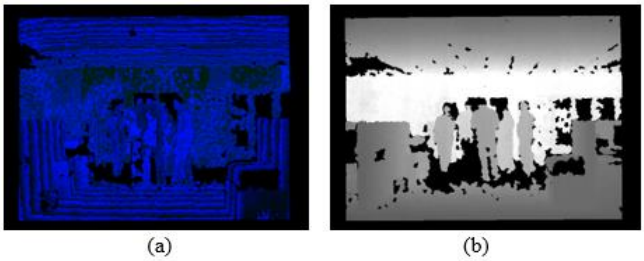

Figure 3. The input image, (a) depth image and (b) thermal image.

Next is to align the *Idpt* using the spatial transformation structure, *dmy_tform* obtained from the control point selection stage so that both images *Idpt* and *tml* are adequately registered for further processing. Figure 4 illustrates the output of this process named *depth_r* images.

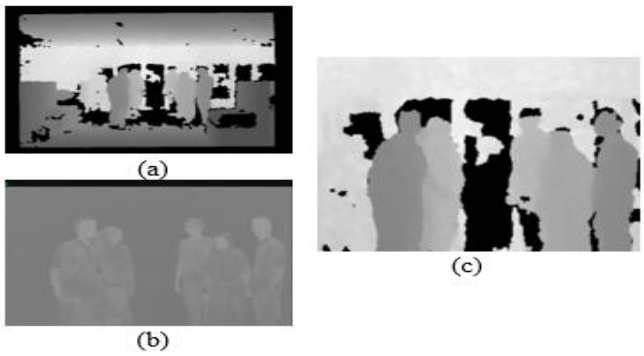

Figure 4. The align process (a) *Idpt* image, (b) thermal image, and (c) *depth_r* image.

3.2. Segmentation Stage

The segmentation of thermal images is based on threshold values chosen from histogram using Image Viewer Application in Image Processing Tools. The Image Viewer provides several capabilities to display images, which optimizes the numbers, axes, and properties of the image objects for image display, as mentioned by [26]. This range of thresholds will be used for segmenting the thermal images. In [15] reported that the thermal images after threshold are named TOI (thermal of interest) in which the test is carried out based on the value of the lower threshold, *tl* and the highest threshold value, *th* using the following formula:

$$TOI(x,y) = \begin{cases} 0: & t_l \geq tml(x,y) \geq t_h \\ tml(x,y): & otherwise \end{cases} \quad (1)$$

where *tml(x, y)* is the original thermal image pixels and *TOI(x, y)* is the thresholded image. If the measured thermal value *tml(x, y)* falls outside the *th* range or falls below the *tl* range, the thermal value is set to zero. Otherwise, the thermal value is maintained. Figure 5 illustrated an example process of segmentation for the thermal image.

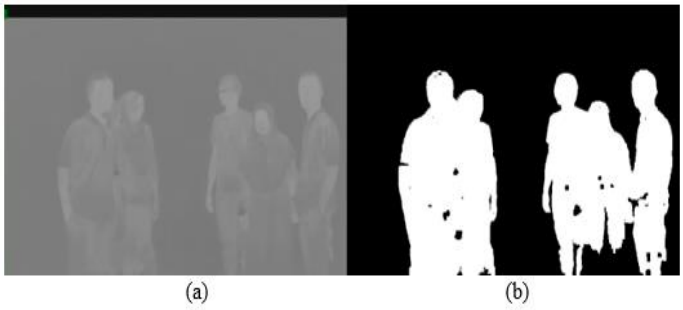

Figure 5. The example process of segmentation for thermal image, (a) thermal image and (b) TOI image.

The segmentation of depth images should be fast and precisely. Thus, here were introduce the new technique of segmentation for depth images named F.T.D.I (**Fusion Thermal-Depth Information**) technique. In general, the Kinect camera is configured to deliver valid range data of about 1.5 to 10 meters [27]. However, only humans in the range of 3 to 8 meters are calculated as valid because the depth range of above 8 meters is inaccurate, while below 3 meters, the thermal image is not suitable due to the focal length of the thermal camera which is 18mm with a horizontal field of view of 29.9°. Therefore, the range of below 3 meters and above 8 meters needs to be eliminated. Next, eliminate the foreground regions that are non-human. This is done by using F.T.D.I technique to generate the depth of interest (*DOI*) with the *TOI* images as a template against *depth_r* images. This technique can be expressed as,

$$Output = A \cdot B \quad (2)$$

$$DOI(x,y) = \begin{cases} 0: & TOI(x,y) \; or \; depth_r(x,y) \; or \; both \; equal \; to \; 0 \\ depth\_r(x,y): & both \; bigger \; than \; 0 \end{cases} \quad (3)$$

where B refers to *depth_r(x, y)* - the depth registered, A refers to *TOI(x, y)* - the thresholded image of thermal and Output refers *to DOI(x, y)* - the thresholded image of depth. Table 1 shows the truth table for the F.T.D.I technique.

Table 1. The truth table for the FTDI technique.

Input		**Output**
TOI (x, y)	***depth_r (x, y)***	***DOI (x, y)***
0	0	0
0	> 0	0
> 0	0	0
> 0	> 0	*depth_r (x, y)*

This truth table with two input images, i.e. *TOI* and *depth_r* images, in which if the pixel values of both images are greater than 0, so the pixel value of *depth_r(x, y)* will be stored to the matrix of *DOI(x, y)* as well as when one of them is zero, the matrix of *DOI(x, y)* becomes zero. These values are based on the pixel value for each input image. Based on the range of detection of 3 to 8 meters, the range of pixel value for *DOI* images is 77 to 255. For instance, here are some sample values for *TOI, depth_r*, and results of DOI, as shown in Table 2.

Table 2. Example function of the truth table.

Input		**Output**
TOI (x, y)	***depth_r (x, y)***	***DOI (x, y)***
0	0	0
0	255	0
128	0	0
228	175	175

where if either the value of *TOI* or *depth_r* or both is zero, the value of *DOI* is zero. Whereas, if the values of *TOI* and *depth_r* are greater than zero (228 and 175), the value of *DOI* is the same value of *depth_r* (175).

4. Experimental and Result

Segmentation of thermal images - to determine the appropriate threshold values for the thermal image segmentation. In addition to that, this experiment was also carried out in both illuminations (real room situation), i.e. dark and bright. Results show that both illuminations do not affect the threshold for obtaining the thermal of interest (*ToI*). To set the threshold values equivalence to the temperature of humans, as many as five images (from the positive dataset) were chosen randomly, in which 3 images from dark illumination and 2 images from bright illumination. The experimental yield shows that 4 different values obtained for *th*, so the final values for *tl* and *th* are chosen, i.e. 135 and 160, respectively.

The segmentation of depth images was not use available techniques because most of the current techniques engage significant computational burden and mathematical complexity usually faced with a problem in real-time implementation. Thus, a new technique of segmentation for depth images is introduced, namely F.T.D.I technique. This proposed technique is performed by comparing every pixel value between depth_r images and *ToI* images, as shown in Table 2. The fusion technique performs well to generate 100% DOI's, as illustrated in the qualitative results in Figure 6.

Evaluation of the performance for the state-of-the-art technique and proposed techniques are conducted using run-time. The total run-time for the proposed technique on the proposed dataset MRV1 and MRV2 (828 frame images) is 23.85sec. Thus, the quantitative results show that the proposed technique performs well with an average run-time of 28.85msec/frame compared to Puji and Schade (2018), and Wanli et al. (2018) with 30.55msec/frame and 31.23msec/frame respectively. Table 3 shows the comparison performance based on run-time between the state-of-the-art techniques and the proposed technique.

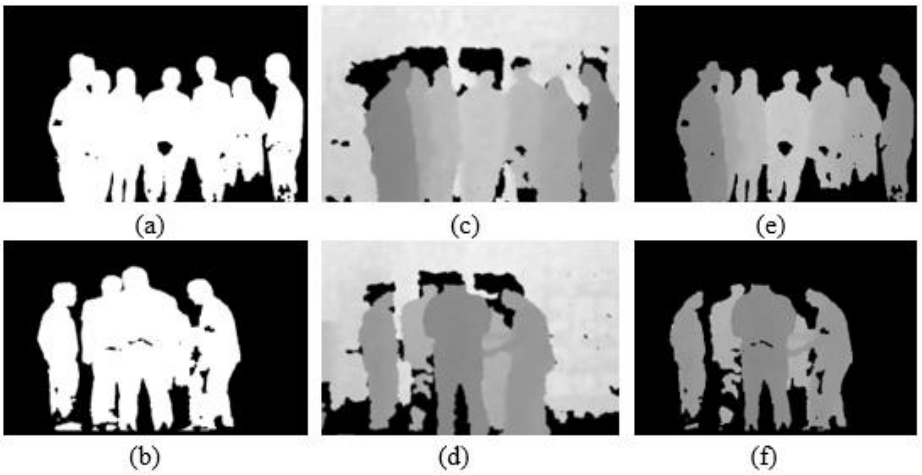

Figure 6. The results of fusion technique, (a-b) thermal of interest, (c-d) depth registered image and (e-f) depth of interest.

Table 3. The comparison performance between proposed technique and state-of-the-art technique using run-time.

Techniques	Total run-time	Average run-time
Proposed Techniques	23.85sec	28.80msec/frames
Puji-Schade Techniques	30.55sec	36.90msec/frame
Wanli Techniques	31.23sec	37.72msec/frame

The expedited processing of segmentation is achieved. This makes it suitable for mobile robots and suitable for operating in real-time. This result achieved the objective of developing a better and faster segmentation technique to produce depth of interest (DOI).

Next evaluation is to know the performance of detection rate based on input images. This experiment is comparing three types of input image (from MRV dataset 828 frames), i.e. thermal, depth and DOI. The total Number of Ground Truth (NGT) is 4296, excluding a person occluded, in which 328 frames with NGT 2296 data for Single Person (SP) and the occluded person has considered false alarms. Table 4 shows the results based on single person dataset.

Table 4. The comparison performance between proposed technique

Type of Images	Detection Metrics			Total Run-Time
	Precision	Recall	Accuracy	
DOI	93.03%	95.07%	94.18%	43.58sec
Thermal	83.78%	87.78%	86.54%	28.11sec
Depth	71.06%	71.51%	74.86%	55.48sec

The quantitative results show that the DOI images (fusion modality) outperform two other input images, in which the rate of accuracy is approximately 94%. Whereas, the accuracy rate for thermal images and depth images is approximately 86% (moderate performance) and 74% (lower performance) respectively. Whereas, the total run-time for depth images is 55.48sec, this is due to the process of segmentation to obtain DOI without additional modality is difficult as well as causes over-segmentation. Whereas, thermal images cannot detect people properly in crowded scenes despite having the fastest processing of segmentation.

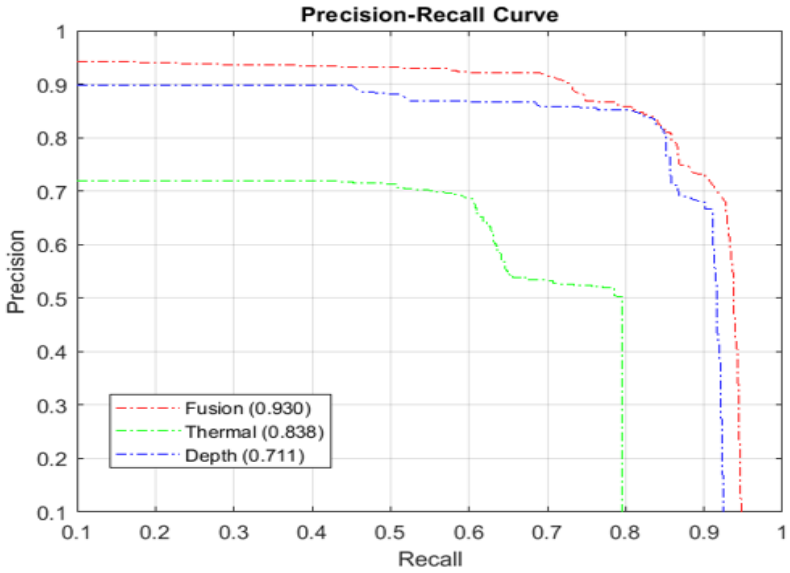

Figure 7. The PRC curve for efficiency of the detection single person.

The qualitative results show the efficiency of the detection single person based on the thermal, depth and DOI images as illustrated in Figure 7. The PRC curve of DOI images (fusion modality) shows that the curve is close to 1, which means that the system using combined modalities has the advantage and precision in human detection versus single modality.

5. Conclusion

The objective has been achieved to develop a better segmentation technique. This is done by proposing F.T.D.I technique to produce the depth-of-interest (DOI). This technique based on information of thermal and depth is ensured to solve computational burden and mathematical complexity, which occurs in many previous segmentation techniques, namely the over-segmentation process. The results show that the system uses the proposed technique to maximize the recall and precision rates, thus achieving a very high accuracy rate. In addition to that, this approach has the versatility for various applications, whether mobile or static platform to distinguish human and not human. The approaches of F.T.D.I technique can help speed up the segmentation process and be able to generate each person's characteristics based on pixel code through the candidate selection process. This proposed technique helps to balance the disadvantages of the two modalities with each other.

References

[1] O. Wanli, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1874–1887, 2018, doi: 10.1109/TPAMI.2017.2738645.

[2] M. Bertozzi, A. Broggi, M. Del Rose, M. Felisa, A. Rakotomamonjy, and F. Suard, "A Pedestrian Detector Using Histograms of Oriented Gradients and a Support Vector Machine Classifier," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2007, pp. 143–148, doi: 10.1109/ITSC.2007.4357692.

[3] T. Gandhi and M. M. Trivedi, "Pedestrian Protection Systems: Issues, Survey, and Challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sep. 2007, doi: 10.1109/TITS.2007.903444.

[4] T. Saba, A. Rehman, Z. Mehmood, H. Kolivand, and M. Sharif, "Image Enhancement and Segmentation Techniques for Detection of Knee Joint Diseases: A Survey," *Curr. Med. Imaging Rev.*, vol. 14, no. 5, pp. 704–715, 2017, doi: 10.2174/1573405613666170912164546.

[5] M. S. Gajjar, Vandit; Khandhediya, Yash; Gurnani, Ayesha; Mavani, Viraj; Raval, "Using Visual Saliency to Improve Human Detection with Convolutional Networks," in *Workshop in Conjunction with CVPR-2018*, 2018, pp. 1–9.

[6] U. J. Reddy, P. Dhanalakshmi, and P. D. K. Reddy, "Image segmentation technique using SVM classifier for detection of medical disorders," *Ing. des Syst. d'Information*, vol. 24, no. 2, pp. 173–176, 2019, doi: 10.18280/isi.240207.
[7] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *arXiv*, pp. 1–58, 2019.
[8] R. Priyadharsini and T. S. Sharmila, "Object Detection in Underwater Acoustic Images Using Edge Based Segmentation Method," *Procedia Comput. Sci.*, vol. 165, pp. 759–765, 2019, doi: 10.1016/j.procs.2020.01.015.
[9] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and and D. Terzopoulos, "Image Segmentation Using Deep Learning : A Survey," *arXiv preprint arXiv:2001.05566*, 2020.
[10] S. Manoharan, "Performance Analysis of Clustering Based Image Segmentation Techniques," vol. 02, no. 01, pp. 14–24, 2020.
[11] S. P. Mary, Ankayarkanni, U. Nandini, Sathyabama, and S. Aravindhan, "A Survey on Image Segmentation Using Deep Learning," *J. Phys. Conf. Ser.*, vol. 1712, no. 1, 2020, doi: 10.1088/1742-6596/1712/1/012016.
[12] X. Zhou, X. Liu, A. Jiang, B. Yan, and C. Yang, "Improving Video Segmentation by Fusing Depth Cues and the Visual Background Extractor (ViBe) Algorithm," *Sensors*, vol. 17, no. 6, p. 1177, 2017, doi: 10.3390/s17051177.
[13] L. Puji and H.-P. Schade, "RGB-Depth Image Based Human Detection Using Viola-Jones and Chan-Vese Active Contour Segmentation," in *International Symposium on Signal Processing and Intelligent Recognition Systems*, no. October 2017, 2018, p. pp 285-296.
[14] L. Wang, K. L. Chan, and G. Wang, "Human Detection with Occlusion Handling by Over-Segmentation and Clustering on Foreground Regions," in *Lecture Notes in Computer Science, ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part II*, 2013, pp. 197–208.
[15] M. S. Kristoffersen, J. V. Dueholm, R. Gade, and T. B. Moeslund, "Pedestrian counting with occlusion handling using stereo thermal cameras," *Sensors (Switzerland)*, vol. 16, no. 1, 2016, doi: 10.3390/s16010062.
[16] G. H. Zhao ZQ., Bian H., Hu D., Cheng W., "Pedestrian Detection Based on Fast R-Cnn and Batch Normalization," in *Intelligent Computing Theories and Application (ICIC 2017), Lecture Notes in Computer Science*, vol. 10361, Springer, Cham, 2017.
[17] M. Antonello, M. Carraro, M. Pierobon, and E. Menegatti, "Fast and Robust Detection of Fallen People from a Mobile Robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 4159–4166, [Online]. Available: http://arxiv.org/abs/1703.03349.
[18] H. S. Hadi, M. Rosbi, U. U. Sheikh, and S. H. M. Amin, "Improved occlusion handling for human detection from mobile robot," 2015, doi: 10.1109/SAI.2015.7237217.
[19] R. Raghavendra and C. Busch, "Novel image fusion scheme based on dependency measure for robust multispectral palmprint recognition," *Pattern Recognit.*, vol. 47, no. 6, pp. 2205–2221, 2014, doi: 10.1016/j.patcog.2013.12.011.
[20] C. Palmero, A. ClapÃ©s, C. Bahnsen, A. MÃ¸gelmose, T. B. Moeslund, and S. Escalera, "Multi-modal RGB–Depth–Thermal Human Body Segmentation," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 217–239, 2016, doi: 10.1007/s11263-016-0901-x.
[21] D. O. Pop, A. Rogozan, F. Nashashibi, and A. Bensrhair, "Pedestrian recognition through different cross-modality deep learning methods," *2017 IEEE Int. Conf. Veh. Electron. Safety, ICVES 2017*, pp. 133–138, 2017, doi: 10.1109/ICVES.2017.7991914.
[22] A. Mateus, D. Ribeiro, P. Miraldo, and J. C. Nascimento, "Efficient and Robust Pedestrian Detection using Deep Learning for Human-Aware Navigation," 2016, [Online]. Available: http://arxiv.org/abs/1607.04441.
[23] T. Luchao, M. Li, Y. Hao, J. Liu, G. Zhang, and Y. Q. Chen, "Robust 3D Human Detection in Complex Environments with Depth Camera," *IEEE Trans. Multimed.*, vol. 20, no. 9, pp. 2249–2261, 2018.
[24] V. Vaquero, M. Villamizar, and A. Sanfeliu, "Real time people detection combining appearance and depth image spaces using boosted random ferns," *Adv. Intell. Syst. Comput.*, vol. 418, pp.

587–598, 2016, doi: 10.1007/978-3-319-27149-1_45.
[25] S. Pourmehr, J. Thomas, J. Bruce, J. Wawerla, and R. Vaughan, "Robust sensor fusion for finding HRI partners in a crowd," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2017, no. i, pp. 3272–3278, doi: 10.1109/ICRA.2017.7989373.
[26] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Image Processing Toolbox User's guide*, vol. 3. 2017.
[27] G. Marchal and T. Lygren, "The Microsoft Kinect : validation of a robust and low-cost 3D scanner for biological science," 2017. doi: 10.13140/RG.2.2.12069.40167.