Perspective independent humanoid object detection by 2D and 3D data analysis

Chengsi Zhang *Electrical and Computer Engineering*
Faculty of Engineering and Applied Science
Memorial University
cz2075@mun.ca

Stephen Czarnuch *Electrical and Computer Engineering*
Faculty of Engineering and Applied Science
Memorial University
sczarnuch@mun.ca

***Abstract*—Identifying moving humans in indoor environments using a camera with unknown orientation is a challenging problem due the pose variation, varying body shapes and potentially complex backgrounds. Existing approaches share common assumptions or conditions, such as long setup time or known camera orientation. We present an approach to segment and recognize humans in an indoor environment given RGB-D data from a camera with arbitrary orientation and location, only assuming that people smoothly move in the camera field of view with their body perpendicular to the ground plane, and that the camera's position remains constant. Methodologically, we identify all moving humanoid objects by evaluating histogram intersection change across video segments, object dimensions and the trajectory vector of the homography decomposition. From a set of 24 RGB-D data trials captured using a Kinect sensor, we identify the largest ground planes, cluster objects in the scenes and find 2D SIFT features for those objects, and then build a motion sequence for each object by evaluating the intersection of each object's histogram in three dimensions across frames. After find the reliable homography for all objects, we identify the moving human object by checking the change in the histogram intersection, object dimensions and the trajectory vector of the homgraphy decomposition. Our results show that the moving humanoid objects can be successfully detected, if visible to both RGB and depth sensor, regardless of camera orientation and movement speed of the human. Overall, our approach robustly estimated the moving humanoid objects in 24 indoor scenarios with arbitrary camera orientation and location.**

***Index Terms*—Human identification, Homography, homography decomposition, point cloud segmentation, 3D data analysis.**

I. Introduction

With one additional dimension, 3D data provide a more intuitive and realistic environmental perspective in computer vision applications than traditional 2D data. By combining traditional 2D RGB data with depth information, 3D data create a more comprehensive digital representation of real world environments, providing considerable value in many applications such as training and simulation [1][2], construction [3][4] and gaming [5][6]. The benefits of 3D data over 2D data are particularly noticeable in cluttered or dynamic environments. In these complex environments, 3D data allow enhanced visual understandings, improved precision and accuracy, easier risk/issue identification and analysis, and intuitive model manipulation [7][8][9]. For example, operating rooms typically have many objects that frequently change depending on the nature of the emergency, including multiple humans who enter and exit the room and interact with the objects and each other. Constructing an accurate 3D model of an operating room and recording videos of various processes within the room could create a helpful and interactive tool for training and simulation, or be used in real time to observe and monitor the room. For applications like gaming, the room is often modified to accommodate placement of a sensor (i.e., clearing out a space), the sensor is intentionally located in an ideal position, and users are willing to undergo a calibration process if necessary. However, the applications we consider, such as the operating room, are complex, dynamic and cluttered real-world environments, where the sensor must be located out of the way of the processes or occupants of the room, and systems using the sensor would need to auto-calibrate because occupants of the room are unlikely to be willing to perform calibrations. Accordingly, in applications in these complex environments, the sensor's location and orientation in the room will generally be unknown (e.g., the sensor's field of view cannot be assumed to be parallel to the ground). In this paper, we focus on addressing the difficulties of segmenting the moving humanoid objects in a indoor environment without any a prior knowledge of the sensor or room, which will be an important prior knowledge to allow future estimation of the unknown orientation of a sensor and relative position between multiple sensors.

Human detection is essential due to its usage in a variety of applications including human tracking[10][11], human identification[12][13], and human statistical processes (e.g., counting human occupancy)[14][15]. Indeed, most computer vision algorithms implicitly assume a known human orientation[16][17] and the visibility of special human features[18][19]. However, in complex and dynamic environments with unknown sensor placement, these assumptions are hard to achieve. In addition, finding moving humanoid objects allows an estimation of the ground plane, and as a result the camera orientation, which then also facilitates improved 3D registration and 3D reconstruction of data from multiple sensors viewing the same scene by converting a 3D problem into a 2D problem. Ultimately our goal is to start by detecting moving humanoid objects, and then estimating the ground plane for each sensor in a multi-sensor system using this human movement, such that the ground plane can be used as a reference for finding the positions and orientations of each sensor relative to each other, which will facilitate the

reliable 3D reconstruction of a medical operation room.
To accomplish our goal, we aim to develop a system that detects moving humanoid objects with unknown positions and orientations in an indoor environment. Our only assumptions are: that most of at least one person can be seen smoothly moving in the RGB-D camera field of view and the RGB-D camera's position and orientation remain unchanged until the humanoid object detection is complete. In order to detect the moving humanoid objects under this condition, we combine the robustness of 3D Random Sample Consensus (RANSAC) and 2D homography decomposition. While 3D RANSAC extracts useful spatial information from each 3D point cloud segment, 2D homography decomposition constructs homography planes from people walking on the ground.

II. Related work

Existing humanoid object detection work can be broadly categorized into 2D or 3D approaches based on the sensor type. Within 2D approaches, the most widespread approach is using Histograms of oriented gradients (HOG)[16], which detects humans based on local feature patterns. Multiple 2D humanoid object detection approaches, such as [17] [20] and [21], are built based on HOG and improve the efficiency and accuracy of HOG by analysing HOG features with Partial Least Squares, combined with Gaussian and Background Contours Subtraction, and using feature approximation respectively. Due to the nature of HOG, all of these approaches are significantly sensitive to the camera orientation, especially to the roll angle. In addition, the approach proposed by Oren Freifeld et al.[22] which used training templates to describe human 2D contour body shape and the color-curve based deep learning neural network method built by Humberto Souto et al.[23] are also fit in to 2D approaches. However, these two approaches require human body parts that are clearly independent from the background. Furthermore, Dragon et al. [24] proposed an approach for detecting and tracking moving objects from RGB images where RGB frames captured from a moving sensor are iteratively split into regions until reliable homographies can be estimated from the feature points within these regions. The decomposition of the homography with the highest probability indicates the orientation and ego motion of the sensor's movement. Unfortunately, this approach is not suitable for indoor environments with a stationary sensor because moving objects will be a small proportion of the scene, making it hard to distinguish between a homography generated from mismatched key points and a homography from a moving object. Further, their solution requires the shape of moving objects to remain unchanged to ensure successful feature correspondence between frames; a condition that cannot be guaranteed in indoor environments with an arbitrary fixed perspective. Although the above 2D approaches can successfully detect human or moving objects, none of them work in dynamic or cluttered environments where the location and orientation of the sensor is unknown.
3D-based Human detection approaches commonly utilize the pattern of 3D local features, which describes some characteristics of 3D raw data, to identify some special components of humans. Rauter proposed a depth map feature descriptor to search and find human head based on the unique depth value distribution of human body-to-shoulder area[25]. Similarly, Xia et al. created a 2D contour model and a 3D surface model to extract the human head from a depth image, and then detected the human body by extending the human head region[19]. Notably, these two algorithms fail to detect a human if the head is not in the camera field of view (FOV) or the shape of head does not fit the pre-designed pattern (e.g. wearing a nurse hat). In addition to the depth map-only approaches, other researchers utilized the combination of RGB and depth data for human detection. The system proposed by Buys et al. combines depth segmentation based on Kinematic modeling[26] and color labeling to detect humans and recognize human poses[27], while Liu et al. detected humans through 3D spatial and 2D RGB data statistics[28]. However, Kinematic models require human body parts that are independent from each other and background, and Liu's approach assumes the human head is always visible. Together, the most robust and reliable 2D and 3D methods of detecting humans in the scene have common assumptions or predicates, such as the known and unchanged orientation of the camera, certain human body parts are visible, or human body parts are independent within the camera's FOV. While these assumptions restrict the complication of the human detection problem based on the requirements of specific applications, they cannot be used in real-world scenarios where the camera location and orientation are unknown, and the environment is complex, cluttered or dynamic. To overcome the limitations of these assumptions for our application, we build on the approach of Dragon et al. [29][24] because the assumptions of their approach are closest to our conditions. Notably, while their approach requires the sensor to be moving, we assume that the sensor is stationary and something in the scene is instead moving. In our case, we will restrict our interest to a human moving in the scene, which is very common in a complex and dynamic scene, though this does not necessarily need to be the case. We present our approach to accomplish this in section III followed by our experimental setup and results in section IV. We then present our discussion and future work in section V.

III. Methodology

Our humanoid object detection approach combines the robustness of 2D and 3D computer vision algorithms. The major components of our approach are: 1) Data pre-processing (section III-A) where we described the preparation of 2D and 3D data with corresponding features; and 2) 2D homography decomposition (section III-B), where we decomposed the 2D homography according to 3D feature restrictions to estimate the trajectory of any moving humanoid objects in the scene.

A. Data pre-processing

To obtain a more useful 3D data representation, we first generated a 3D point cloud from the RGB-D data using the

intrinsic and extrinsic parameters of the sensor. We calibrated using Zhang's approach with the intrinsic parameter matrix defined as: [30]:

$$K_c = \begin{bmatrix} f \times m_x & \gamma & u_0 \\ 0 & f \times m_y & v_0 \\ 0 & 0 & 1 \end{bmatrix},$$

where f is the focal length, m is the scale factor, γ is the skew coefficient between the x and y axes, and (u_0, v_0) is the principal point. The extrinsic parameter matrix is $\begin{bmatrix} R_{3\times3} & T_{3\times1} \\ 0_{1\times3} & 1 \end{bmatrix}$, composed of rotation and translation parameters R and T. Finally, using radial distortion k_1, k_2, k_3 and tangential distortion p_1, p_2 coefficients, we calculated the camera matrix C by multiplying the intrinsic and extrinsic matrices, such that the depth images were undistorted based on camera parameters and distortion coefficients [31] according to

$$\begin{aligned} X &= p_2(3x^2 + y^2) + x(k_2(x^2 + y^2) \\ &+ k_1(x^2 + y^2) + 1) + 2P_1xy \quad (1) \\ Y &= p_1(x^2 + 3y^2) + y(k_2(x^2 + y^2)^2 \\ &+ k_1(x^2 + y^2) + 1) + 2p_2xy \quad (2) \\ Z &= z \quad (3) \end{aligned}$$

From Eqs.(1), (2) and (3), the coordinates and value of each pixel in each depth image was transformed to an individual point in the associated 3D point cloud.

In general, the point cloud of an indoor environment is composed of planes (e.g., walls, floor), objects (e.g, drawers, chairs), and humans, though in some cases substantial portions of objects are also planes (e.g., desks). In a cluttered indoor environment, planes normally occupy the majority part of the camera FOV. Therefore, after down-sampling the point cloud by applying a voxel grid filter, we disassembled each point cloud into plane segments and object segments by storing and removing any planes greater than 20% of the original point cloud, and then clustering the remaining objects. We used Random Sample Consensus(RANSAC)[32] to extract the planes.

After we stored and removed the planes in the scene, we segmented the point cloud into object clusters using Euclidean clustering algorithm[33]. We first employed Euclidean clustering to find groups of points that were physically close to each other, and then we stored all clustered objects S_o and extracted planes S_p.

To identify which clustered objects are moving in the scene in preparation for homography estimation, we needed to find corresponding objects between successive frames. We utilized Scale-Invariant Feature Transform (SIFT)[34] as the feature extractor on the RGB images to derive 2D feature points. SIFT was able to generate a sufficient number of 2D features for each object in the scenes; particularly for any humans. Additionally, SIFT accommodates a wide range of performance control through variation of the octave layer number $nOct$, edge-like feature filter threshold $eThresh$, and the sigma of Gaussian filter σ [35], allowing excellent optimization for keypoint detection. For each RGB frame, the 2D feature points were stored as an output of the data preparation phase, along with the 3D points of the clustered objects and the extracted planes.

B. Homography estimation

We used the homography [36] between moving objects across successive frames to construct a plane representing the orientation of moving objects. With a minimal sample set of four feature key point correspondences between frames at time t and time $t + \Delta t$, a nine-parameter homography matrix H:

$$H = \begin{bmatrix} h_1 1 & h_1 2 & h_1 3 \\ h_2 1 & h_2 2 & h_2 3 \\ h_3 1 & h_3 2 & h_3 3 \end{bmatrix}$$

can be generated, which represents the transformation between 2D points in image coordinates and 3D points in the camera coordinate system.

To find which objects were moving between successive frames, we implemented the Blockwise Linearity Assumption (see [29]). We segmented the entire data set into blocks $B = \{F_1, F_2, ...F_x\}$ of frames F ranging from frame 1 to x. Let S_o^1 and S_o^2 denote all the object segments in the first and second point clouds representing a pair of successive frames. We calculated the 1-D histogram of three dimensions $Hist_x$, $Hist_y$, $Hist_z$ for each object segment $S_{o_i}^1$ and $S_{o_j}^2$. Then, we matched a pair of object segments in F_1 and F_x that represented the same object O_i by determining if the intersection ratio

$$intersection_{o_i}^x = \frac{A(S_{o_i}^1) \cap A(S_{o_j}^x)}{A(S_{o_i}^1)} \quad (4)$$

between the histogram areas of $S_{o_i}^1$ and $S_{o_j}^2$ was greater than zero, and decreased as x increased. To ensure the histogram intersection was larger than zero between the first frame F_1 and frame F_x, we chose a small block size similar to [24][29]. The resulting list of matched pairs of 3D objects $S_{o_i}^1$ and $S_{o_j}^2$, including any moving humans, were projected to 2D pixel clusters $C_{o_i}^1$ and $C_{o_j}^2$ according to

$$\begin{aligned} x_2 &= x_1(1 + k_1r^2 + k_2r^4) + 2p_1x1y1 \\ &+ p_2(r^2 + 2x_1^2) \quad (5) \\ y_2 &= y_1(1 + k_1r^2 + k_2r^4) + 2p_1x1y1 \\ &+ p_2(r^2 + 2y_1^2), \quad (6) \end{aligned}$$

where (x_2, y_2) denotes the distorted pixel coordinates and $r = \sqrt{x_1^2 + y_1^2}$. Consequently, each 2D pixel cluster $C_{o_i}^x$ is then converted to a 2D feature point cluster $R_{o_i}^x$ by using each 2D pixel (x_i, y_i) as the center point and searching for the closest feature points within the radius τ.

We removed any feature keypoints that were outside of the regions, and applied Motion-Split-And-Merge (MSAM)[24] to each pair of corresponding regions $R_{o_i}^1$ and $R_{o_j}^x$ in F_1 and F_x respectively to find the most reliable keypoint clusters for generating a homography $H_{o_i}^x$. We then decomposed each homography $H_{o_i}^x$ into the four plane normal vector,

trajectory vector, and rotation vector solutions $D^x_{o_i{}^{1\sim4}} = \{\vec{n}^i_{o_i{}^{1\sim4}}, \vec{t}^i_{o_i{}^{1\sim4}}, \vec{r}^i_{o_i{}^{1\sim4}}\}$[37], and filtered out the invalid solutions to construct the most reliable decomposition solution $B = \{\vec{n}_{o_i}, \vec{t}_{o_i}, \vec{r}_{o_i}\}$ for each 2D object region R_{o_i} within a block. Here, invalid homography solutions were characterized by checking if a key point (x_i, y_i) within region R_{o_j}, which yields $\vec{z_i} < 0$ $\big((\vec{x_i}, \vec{y_i}, \vec{z_i}) = H \times (x_i, y_i) \quad \wedge \quad \vec{n}^T_{o_i} \times (\vec{x_i}, \vec{y_i}, \vec{z_i}) = 1\big)$, exists[29]. Finally, we built the set of all the moving objects in the scene S_{mo_i} by extracting the object regions that had large and successively decreasing differences in intersection coefficient $intersection^x_{o_i}$ among all objects O in a block. According to the assumption that the person body is perpendicular to the ground plane while moving, we use a cascade filter, which includes the longest edge E_l of moving object bounding boxes larger than a length threshold $Thresh_l$; the ratios between the longest edge E_l and other two edges are larger than a ratio threshold $Thresh_r$; and trajectory vector $\vec{t}^i_{o_i}$ is perpendicular to the longest edge of object bounding box E_l, to determine the moving humanoid object among all moving objects.

IV. Experiments

We collected video sequence data using the Kinect v1 which provides an RGB image and a depth image with a 27 frame per second rate (FPS) on average, image data we combine to form an RGB-D image, using a MacBook Pro (Retina, 13-inch, Mid 2014) with Dual core i5 CPU and 8G memory. We recorded video sequences by placing the camera in 24 unique scenarios, which included various combinations of different camera orientations and locations, multiple planes, multiple people, diverse moving speeds, and various body appearance ratios.

Our captured video sequences contained 40-140 data frames from the time the first person entered the camera's field of view or started moving to the time the last person left the camera FOV or stopped moving. From experimentation, we determined that planes with a confidence score $\zeta > 8.5$ are highly likely to be the actual ground plane, while planes with a confidence score of $6.0 < \zeta < 8.5$ are planes that are parallel to the ground plane, and may be the ground plane. Figures 1 and 2 show examples of our human detection results where human are marked with green dots.

In the data preparation step, SIFT generated an average of approximately 4,000 keypoints in each full 2D image with 10 layers in each octave, 0.02 as contrast threshold, 20 as filter out edge-like features threshold, and 1.0 as sigma. The size of voxel grid down-sample filter for point cloud frames we selected was 2cm. The RANSAC distance threshold and the cluster tolerance of Euclidean clustering were 2.5 and 2 times the voxel grid filter size respectively. Based on these parameters, we extracted anywhere from 4 to 10 planes from each scene, varying based on the indoor environment complexity and camera perspective. In the homography estimation step (section III-B), we set the block size to five to ensure we achieved sufficient histogram intersection between the reference frame F_1 and frame F_x. The number of SIFT feature keypoints on the human ranged from 150 to 380 out of the approximately 4,000 keypoints.

V. Discussion and Conclusions

In this paper, we proposed a novel human detection method using the combination of 2D and 3D data analyses. Our approach helps us to find the human with only four unrestrictive assumptions: the sensors is an RGB-D camera; a person smoothly walks in the scene with most parts of the body visible within the camera field of view; the human body is perpendicular to the ground plane while walking; and the RGB-D camera position and orientation remain unchanged until ground plane estimation is complete. Our approach robustly detects the human with a large variety of sensor orientations and different room complexities. Our experimental results show that our algorithm is insensitive to the movement speed of walking humans and is tolerant to partial occlusion of the human body. In all cases, we were able to detect the human using RGB-D sensors data without any pre-calibration.

In the future, we will focus on improving the performance of the algorithm; switching to a better RGB-D sensors which provides higher quality data; enhancing the robustness and accuracy of the human object detection algorithm; and achieving potential human recognition or identification within a RGB-D camera system. In addition, we will also test our algorithm on video sequences that have higher indoor complexity and more people visible in the scene.

References

[1] V. Waran, V. Narayanan, R. Karuppiah, D. Pancharatnam, H. Chandran, R. Raman, Z. A. A. Rahman, S. L. Owen, and T. Z. Aziz, "Injecting realism in surgical training—initial simulation experience with custom 3d models," *Journal of surgical education*, vol. 71, no. 2, pp. 193–197, 2014.

[2] A. G. Bruzzone and F. Longo, "3d simulation as training tool in container terminals: The trainports simulator," *Journal of Manufacturing Systems*, vol. 32, no. 1, pp. 85–98, 2013.

[3] R. Sacks, C. M. Eastman, and G. Lee, "Parametric 3d modeling in building construction with examples from precast concrete," *Automation in construction*, vol. 13, no. 3, pp. 291–312, 2004.

[4] Y. W. D. Tay, B. Panda, S. C. Paul, N. A. Noor Mohamed, M. J. Tan, and K. F. Leong, "3d printing trends in building and construction industry: a review," *Virtual and Physical Prototyping*, vol. 12, no. 3, pp. 261–276, 2017.

[5] K. Adams, "3d enhancements to gaming components in gaming systems with real-world physics," Oct. 30 2018, uS Patent App. 10/115,261.

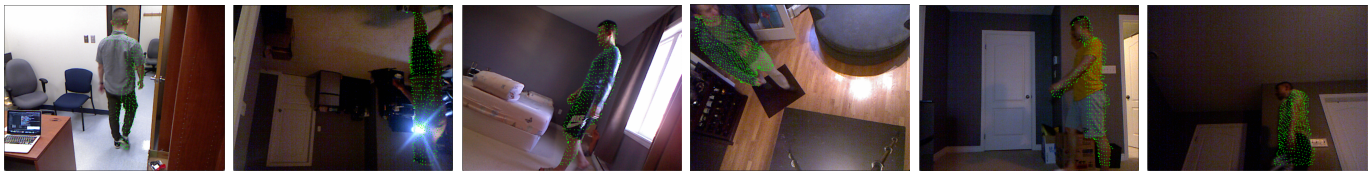
[6] A. Kulshreshth, K. Pfeil, and J. J. LaViola, "Enhancing the gaming experience using 3d spatial user interface technologies," *IEEE computer graphics and applications*, vol. 37, no. 3, pp. 16–23, 2017.

[7] D. M. Swanson, "Benefits of 3d breast tomosynthesis combined with 2d digital mammography in screening women for breast cancer," 2019.

[8] Y. Kudo and N. Ikeda, "Benefits of lung modeling by high-quality three-dimensional computed tomography for thoracoscopic surgery," *Video-Assisted Thoracic Surgery*, vol. 4, 2019.

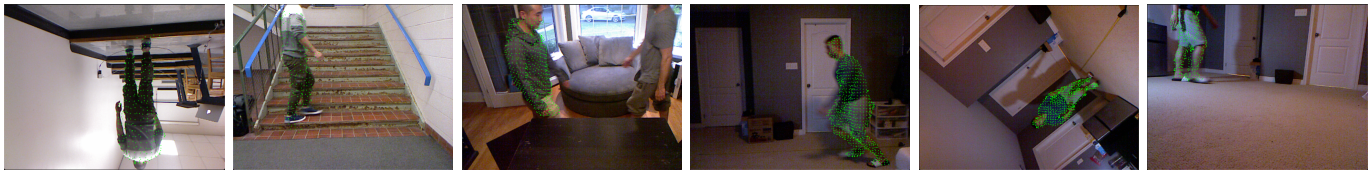
[9] S. J. Trenfield, A. Awad, A. Goyanes, S. Gaisford, and A. W. Basit, "3d printing pharmaceuticals: drug development to frontline care," *Trends in pharmacological sciences*, vol. 39, no. 5, pp. 440–451, 2018.

[10] A. Farooq, A. Jalal, and S. Kamal, "Dense rgb-d map-based human tracking and activity recognition using skin joints features and self-organizing map." *KSII Transactions on Internet & Information Systems*, vol. 9, no. 5, 2015.



(a) lab 0°roll angle away (b) bedroom 180 roll angle 45°pitch (c) bedroom 30°roll angle (d) room -80°pitch angle (e) bedroom 0°pitch angle (f) bedroom 45°pitch angle

Fig. 1: Moving humanoid object detection result of various camera orientations



(a) Multiple large planes (b) Multiple small planes (c) 2 people different direction (d) Moving speed: 2.6 m/s (e) Moving speed: 3 m/s with rotation (f) Small body appearance

Fig. 2: Moving humanoid object detection result of different special scenarios

[11] S. Kamal and A. Jalal, "A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors," *Arabian Journal for science and engineering*, vol. 41, no. 3, pp. 1043–1051, 2016.

[12] F. Okura, S. Ikuma, Y. Makihara, D. Muramatsu, K. Nakada, and Y. Yagi, "Rgb-d video-based individual identification of dairy cows using gait and texture analyses," *Computers and Electronics in Agriculture*, vol. 165, p. 104944, 2019.

[13] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5380–5389.

[14] D. Liciotti, A. Cenci, E. Frontoni, A. Mancini, and P. Zingaretti, "An intelligent rgb-d video system for bus passenger counting," in *International Conference on Intelligent Autonomous Systems*. Springer, 2016, pp. 473–484.

[15] D. Song, Y. Qiao, and A. Corbetta, "Depth driven people counting using deep region proposal network," in *2017 IEEE International Conference on Information and Automation (ICIA)*. IEEE, 2017, pp. 416–421.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[17] A. H. Ahmed, K. Kpalma, and A. O. Guedi, "Human detection using hog-svm, mixture of gaussian and background contours subtraction," in *2017 13th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, Dec 2017, pp. 334–338.

[18] F. Su, G. Fang, and J. J. Zou, "Human detection using a combination of face, head and shoulder detectors," in *2016 IEEE Region 10 Conference (TENCON)*, Nov 2016, pp. 842–845.

[19] L. Xia, C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *CVPR 2011 WORKSHOPS*, June 2011, pp. 15–22.

[20] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 24–31.

[21] D. Sangeetha and P. Deepa, "Efficient scale invariant human detection using histogram of oriented gradients for iot services," in *2017 30th International Conference on VLSI Design and 2017 16th International Conference on Embedded Systems (VLSID)*. IEEE, 2017, pp. 61–66.

[22] O. Freifeld, A. Weiss, S. Zuffi, and M. J. Black, "Contour people: A parameterized model of 2d articulated human shape," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 639–646.

[23] H. Souto and S. Musse, "Automatic detection of 2d human postures based on single images," in *2011 24th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2011, pp. 48–55.

[24] R. Dragon, B. Rosenhahn, and J. Ostermann, "Multi-scale clustering of frame-to-frame correspondences for motion segmentation," in *European Conference on Computer Vision*. Springer, 2012, pp. 445–458.

[25] M. Rauter, "Reliable human detection and tracking in top-view depth images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 529–534.

[26] W. Khalil, E. Dombre, and M. Nagurka, "Modeling, identification and control of robots," 2003.

[27] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter, and C. Pantofaru, "An adaptable system for rgb-d based human body detection and pose estimation," *Journal of visual communication and image representation*, vol. 25, no. 1, pp. 39–52, 2014.

[28] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen, "Detecting and tracking people in real time with rgb-d camera," *Pattern Recognition Letters*, vol. 53, pp. 16–23, 2015.

[29] R. Dragon and L. Van Gool, "Ground plane estimation using a hidden markov model," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Conference Proceedings, p. 4026–4033.

[30] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, 2000.

[31] J. Heikkila, O. Silven *et al.*, "A four-step camera calibration procedure with implicit image correction," in *cvpr*, vol. 97, 1997, p. 1106.

[32] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[33] R. Rusu, "Semantic 3d object maps for everyday manipulation in human living environments," *KI - Künstliche Intelligenz*, vol. 24, no. 4, pp. 345–348, 2010.

[34] D. G. Lowe *et al.*, "Object recognition from local scale-invariant features." in *iccv*, vol. 99, no. 2, 1999, pp. 1150–1157.

[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[36] A. Criminisi, I. Reid, and A. Zisserman, "A plane measuring device," *Image and Vision Computing*, vol. 17, no. 8, pp. 625–634, 1999. [Online]. Available: http://www.robots.ox.ac.uk/ vgg

[37] Y. Ma, S. Soatto, J. Kosecká, and S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*, ser. Interdisciplinary Applied Mathematics. Springer New York, 2005. [Online]. Available: https://books.google.ca/books?id=cbX6jAZ_tggC