People Detection and Tracking from an RGB-D Camera in Top-View Configuration: Review of Challenges and Applications

Daniele Liciotti, Marina Paolanti(✉), Emanuele Frontoni, and Primo Zingaretti

Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche,
Via Brecce Bianche, 60131 Ancona, Italy
{d.liciotti,m.paolanti}@pm.univpm.it, {e.frontoni,p.zingaretti}@univpm.it

Abstract. This paper presents a literature review on the use of RGB-D camera for people detection and tracking. Our aim is to use this state-of-the-art report to demonstrate the potential of top-view configuration for people detection and tracking applications in several sub-domains, to outline key limitations and to indicate areas of technology, where solutions for remaining challenges may be found. The survey examines the success of RGB-D cameras because of their affordability and for the additional rough depth information coupled with visual images that provide. These cameras in configuration top-view have already been successfully applied in the several fields to univocally identify people and to analyse behaviours and interactions. From this report, it emerges that detecting and tracking people can be a valuable source of information for many fields and purposes.

Keywords: Top-View · RGB-D camera · Tracking · Detection

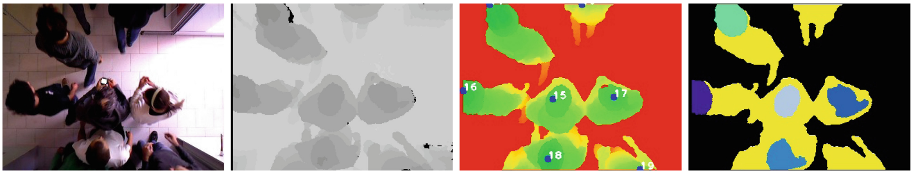
1 Introduction

Detecting and tracking people is an important and fundamental component for many interactive and intelligent systems. The problem remains largely open due to several serious challenges, such as occlusion, change of appearance, complex and dynamic background [26]. Popular sensors for this task are RGB-D cameras because of their availability, reliability and affordability. Studies have demonstrated the great value (both in accuracy and efficiency) of depth camera in coping with severe occlusions among humans and complex background. The appearance of devices, such as Microsoft's Kinect and Asus's Xtion Pro Live Sensors motivates a revolution in computer vision and vision related research. The combination of high-resolution depth and visual information opens up new challenges and opportunities for activity recognition and people tracking for many application fields that ranges from retail to Ambient Assisted Living (AAL). Reliable depth maps can provide valuable additional information to significantly improve tracking and detection results.

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 207–218, 2017.
https://doi.org/10.1007/978-3-319-70742-6_20

The task of detecting and tracking people in such image and sequences has proven very challenging although sustained research over many years has created a range of smart methods. Techniques involve extracting spatially global features and using statistical learning with local features and boosting, such as EOH [10], HOG [7] and edgelet [37]. Other challenges such as high variation in human poses, self-occlusions and cross-occlusions make the problem even more complicated.

To counter these challenges, several research papers adopt the top-view configuration because it eases the task and makes simple to extract different trajectory features. This setup also introduces robustness, due to the lack of occlusions among individuals. Figure 1 depicts a people counting system from top-view configuration with an RGB-D camera.



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

Fig. 1. People counting system from top-view configuration with RGB-D camera. (Color figure online)

The objective of this survey is to provide a comprehensive overview of the recent developments of people detection and tracking with RGB-D technologies from the perspective of top-view, mainly published in the computer vision and machine intelligence communities. The criteria for topic selection arises from our previous experience with approaches with RGB-D cameras installed in a top-view configuration. More specifically, the review includes person tracking and recognition, human activity analysis, hand gesture recognition, and fall detection in different fields. The broad diversity of topics clearly shows the potential impact of top-view configuration in computer vision. We also summarize main paths that most approaches follow and point out their contributions. We categorize and compare the reviewed approaches from multiple perspectives, including information modality, representation coding, structure and transition, and feature engineering methodology, and analyse the pros and cons of each category.

The rest of the paper is organized as follows: Sect. 2 is an overview of the research status on algorithms and approaches adopted with RGB-D sensors installed in a top-view configuration; Sect. 3 introduces the main research fields in which these sensors are installed and employed, such as Security and Video Analytics (Subsect. 3.1), Intelligent Retail Environment (Subsect. 3.2) and Activities of Daily Living (ADLs) (Subsect. 3.3); final section presents the conclusions and our considerations (Sect. 4).

2 Top-View Configuration: Algorithms and Approaches

Many vision techniques and algorithms for person detection and tracking have been proposed during the last years and these greatly restrict the generality of the approach in real-world settings. In this section, we survey current methods, covering both early and recent literature related to algorithms and techniques applied for tracking and detecting humans from top view RGB-D data. In particular, we review the approaches related to segmentation using background subtraction, Water Filling and Statistical algorithms.

Kouno et al. in [17] describe an image-based person identification task focusing on an image from an overhead camera. The process is based on the background subtraction approach. They apply four features to the identification method, i.e. estimated body height, estimated body dimensions, estimated body size and depth histogram.

In [38], the authors propose a system for passengers counting in buses based on stereovision. The processing chain corresponding to this counting system involves different steps dedicated to the detection, segmentation, tracking and counting. In fact, they have segmented the height maps for highlighting the passengers heads at different levels (i.e. adults, teenagers, children). The result is binary images that contain information related to the heads, called "kernels". The extraction part attributes a number of parameters to the kernel such as, size of the kernel, shape, average greylevel, average height level. Then, with the kernel information, a tracking procedure is applied to analyse the trajectories of the kernels.

The top-view camera setting is also adopted in [25]. In this paper, each depth image in a sequence is segmented into K layers as the computer tomography (CT) slides where the depth spacing between two adjacent layers is set to be a fixed value, distance and the number K is an a priori chosen parameter. After that, the region of each slide can be found based on the classic contour finding algorithm. Dynamic time warping algorithm is also applied to address the different sequence length problem. Finally, a SVM classifier is trained to classify the activities.

In another work the authors with methods of low-level segmentation and tracking develop a system that maps the customers in the store, detects the interactions with products on the shelves and the movement of groups of people within the store [24].

Microsoft Kinect depth sensor is employed in [12] in an "on-ceiling" configuration based on the analysis of depth frames. The elements acquired in the depth scene are recognized by a segmentation algorithm, which analyses the raw depth data directly provided by the sensor. The system extracts the elements, and implements a solution to classify all the blobs in the scene. Anthropometric relationships and features are used to recognize human subjects among the blobs. Once a person is detected, he is followed by a tracking algorithm between different frames.

Dittrich et al. [9] present an approach for low-level body part segmentation based on RGB-D data. The RGB-D sensor is installed at the ceiling and observed a shared workspace for human-robot collaboration in the industrial domain.

The object classes are the distinct human body parts: Head, Upper Body, Upper and Lower Arm, Hand, Legs and the background rejection. For the generation of data for the classifier training, they use a synthetic representation of the human body in a virtual environment, where synthetic sensors generate depth data.

A variant of classical segmentation is the one proposed by Tseng in [36]. In this paper, they present a real-time indoor surveillance system which installs multiple depth cameras from vertical top-view to track humans. The system with a framework tries to solve the traditional challenge of surveillance through tracking of multiple persons, such as severe occlusion, similar appearance, illumination changes, and outline deformation. The background subtraction of the stitched top-view image has been performed to extract the foreground objects in the cluttered environment. The detection scheme involves different phases such as the graph-based segmentation, the head hemiellipsoid model, and the geodesic distance map. Furthermore, the shape feature based on diffusion distance has been designed to verify the human tracking hypotheses within particle filter.

An improvement of the classical segmentation techniques is the algorithm proposed by Kepski et al. [15]. The first step of the algorithm is nearest neighbor interpolation to fill the holes in the depth map and to get the map with meaningful values for all pixels. Then, the median filter with a 5 × 5 window on the depth array is executed to make the data smooth. The algorithm also extracts the floor and removes their corresponding pixels from the depth map. Given the extracted person in the last depth frame, the region growing is performed to delineate the person in the current frame. To confirm the presence of the tracked subject as well as to give head location a Support Vector Machine (SVM) based person finder is used. On the basis of the persons centroid the pan-tilt head rotates the camera to keep the head in the central part of the depth map. Finally, a cascade classifier consisting of lying pose detector and dynamic transition detector is carried out.

An additional paper that describes a method for people counting in public transportation with a segmentation approach is [28]. Kinect sensor mounted vertically has been employed to acquire an images database of 1–5 persons, with and without body poses of holding a handrail. However, in this case the image is processed in blocks in order to find potential local maxima, which are subsequently verified to find head candidates. Finally, non-head objects have been filtered out, based on the ratio of pixels with similar and near-zero value, in the neighbourhood of the maxima.

The approach in [3] investigated a real time people tracking system able to work even under severe low-lighting conditions. The system relies on a novel active sensor that provides brightness and depth images based on a Time of Flight (TOF) technology. This is performed by means of a simple background subtraction procedure based on a pixelwise parametric statistical model. The tracking algorithm is efficient, being based on geometrical constraints and invariants. Experiments are performed under changing lighting conditions and involving multiple people closely interacting with each other.

The same technique is the one applied in [39]. In this paper, the method is composed by two behaviour estimators. The first one is based on height of hand with depth information the second instead on SVM with depth and PSA (Pixel State Analysis) based features and these estimators are used by cascading them.

A method to detect human body parts in depth images based on an active learning strategy is proposed in [4]. The approach is evaluated on two different scenarios: the detection of human heads of people lying in a bed and the detection of human heads from a ceiling camera. The proposal is to reduce both the training processing time and the image labelling efforts, combining an online decision tree learning procedure that is able to train the model incrementally and a data sampling strategy that selects relevant samples for labelling The data are grouped into clusters using as features the depth pixel values, with an algorithm such as k-means.

Tian et al., in [35] have adopted the median filtering to noise removal, because it could well filter the depth image noise obtained by Kinect, and at the same time could protect edge information well. A human detection method using HOG features, that are local descriptors, of head and shoulder based on depth map and detecting moving objects in particular scene is used. SVM classifier has isolated regions of interest (features of head and shoulder) to achieve real-time detection of objects (pedestrian).

A method for human detection and tracking in depth images captured by a top-view camera system is presented in [32]. They have introduced feature descriptor to train a head-shoulder detector using a discriminative class scheme. A separate processing step has ensured that only a minimal but sufficient number of head-shoulder candidates is evaluated. A final tracking step reliably propagated detections in time and provides stable tracking results. The quality of the method has allowed to recognise many challenging situations with humans tailgating and piggybacking.

An interesting binary segmentation approach is the one proposed by Wu et al. [37] that have used a Gaussian Mixture Models algorithm and reduced depth-sensing noise from the camera and background subtraction. Moreover, the authors have smoothed the foreground depth map using a 5 by 5 median filter. The real-time segmentation of a tracked person and their body parts has been the first phase of the EagleSense tracking pipeline.

In [13] the authors described and evaluated a vision-based technique for tracking many people with a network of stereo camera sensors. They have modelled the stereo depth estimation error as Gaussian and track the features using a Kalman filter. The feature tracking component starts by identifying good features to track using the Harris corner detector. It has tracked the left and right image features independently in the time domain using Lucas-Kanade-Tomasi (LKT) feature tracking. The approach has been evaluated using the MOTA-MOTP multi-target tracking performance metrics on real data sets with up to 6 people and on challenging simulations of crowds of up to 25 people with uniform appearance. This technique uses a separate particle filter to track each person and thus a data association step is required to assign 3D feature measurements to individual trackers.

Migniot in papers such as [30,31] has addressed the problem of the tracking of 3D human body pose from depth image sequences given by a Xtion Pro-Live camera. Human body poses have been estimated through model fitting using dense correspondences between depth data and an articulated human model. Two trackers using particle filter have been presented.

A computer vision algorithm adopted by many researchers in case of RGB-D cameras placed in top-view configuration is Water filling.

Zhang et al. [40] have built a system with vertical Kinect sensor for people counting, where the depth information is used to remove the effect of the appearance variation. Since the head is closer to the Kinect sensor than other parts of the body, people counting task found the suitable local minimum regions. The unsupervised water filling method finds these regions with the property of robustness, locality and scale-invariance.

Even in [1] and in [6], the authors have presented a water filling people counting algorithm using depth images acquired from a Kinect camera that is installed vertically, i.e., pointing toward the floor. The algorithm in [1] is referred to as Field seeding algorithm. The people head blobs are detected from the binary images generated with regard to the threshold values derived from the local minimum values. In [6] the approach called as people tracking increases the performance of the people counting system.

3 Top-View Configuration: Challenges and Opportunities in Research Fields

The main motivating factors for the installation of RGB-D cameras in top-view configuration are brought back to some related applications that we describe in this section. Firstly, the reliable and occlusion free counting of persons that is crucial to many applications. Most previous works can only count moving people from a single camera, which cannot count still people or can fail badly when there is a crowd and occlusions are very frequent. In this survey, we have focused on the works with RGB-D cameras in top-view configuration in three fields of research: Security and Video Analytics, Intelligent Retail Environment and ADLs.

3.1 Security and Video Analytics

The applications developed in this field are related to safety and security in crowded environments, people flow analysis and access control as well as counting. Actual tracking accuracy of top-view cameras over-performs all other point of view in crowded environments with accuracies up to 99%. When there are special security applications or the system is working in usually crowded scenarios the proposed architecture and point of view is the only suitable.

In [5], the authors have focused on the development of an embedded smart camera network dedicated to track and count people in public spaces. In the network, each node is capable of sensing, tracking and counting people while communicating with the adjacent nodes of the network. Each node uses a 3D-sensing camera positioned in a downward-view. This system has performed background modelling during the calibration process, using a fast and lightweight segmentation algorithm.

A vision based method for counting the number of persons which cross a virtual line is presented in [8]. The method analyses the video stream acquired by a camera mounted in a zenithal position with respect to the counting line, allowing to determine the number of people that cross the virtual line and providing the crossing direction for each person. This approach was designed to achieve high accuracy and computational efficiency. An extensive evaluation of the method has been carried out taking into account the factors that may impact on the counting performance and, in particular, the acquisition technology (traditional RGB camera and depth sensor), the installation scenario (indoor and outdoor), the density of the people flow (isolated people and groups of persons), the acquisition frame rate, and the image resolution. They also analysed the combination of the outputs obtained from the RGB and depth sensors as a way to improve the counting performance.

Another work for people counting is done in [11]. An algorithm by multimodal joint information processing for crowd counting is developed. In this method, the authors have used colour and depth information together with an ordinary depth camera (e.g. Microsoft Kinect). Firstly, they have detected each head of the passing or still person in the surveillance region with adaptive modulation ability to varying scenes on depth information. Then, they have tracked and counted each detected head on colour information.

In order to guarantee security in e.g. critical infrastructure a pipeline is presented in [34]. It verifies that only a single, authorized subject can enter a secured area. Verification scenarios are carried out by using a set of RGB-D images. Features, invariant to rotation and pose are used and classified by different metrics to be applied in real-time.

The combination of the people counting problem with re-identification and trajectory analysis is faced in [14]. They have extracted useful information using depth cameras. The re-identification task is studied by [27]. The authors have introduced a study on the use of different features exclusively obtained from depth images captured with top-view RGB-D cameras. TVPR is the dataset for person re-identification with an RGB-D camera in a top-view configuration. The registrations are made in an indoor scenario, where people pass under the camera installed on the ceiling [23].

3.2 Intelligent Retail Environment

An important scope is the interactions detection between people and environment with the many applications for the field of Intelligent retail environment and intelligent shelf such as Shopper Analytics [18]. The aim of this paper is to present a low cost integrated system consisting of a RGB-D camera and a software able to monitor shoppers. The camera installed in above the shelf detects the presence of people and univocally identifies them. Through the depth frames, the system detects the interactions of the shoppers with the products on the shelf and determines if a product is picked up or if the product is taken and then put back and finally, if there is not contact with the products.

The same authors, in [20] have described the monitoring of consumer behaviours. The autonomous and low cost system employed is based on a software infrastructure connected to a video sensor network, with a set of computer vision algorithms, embedded in the distributed RGB-D cameras.

GroupTogether is another system that explores cross-device interaction using two sociological constructs [29]. It supports fluid, minimally disruptive techniques for co-located collaboration by leveraging the proxemics of people as well as the proxemics of devices.

Migniot et al. have explored the problem of people tracking with a robust and reliable markerless camera tracking system for outdoor augmented reality using only a mobile handheld camera. The method was particularly efficient for partially known 3D scenes where only an incomplete 3D model of the outdoor environment was available [31].

3.3 Activities of Daily Living (ADLs)

Another research field with RGB-D camera top-view is ADLs. In this field the application range goes from high reliability fall detection to occlusion free Human Behaviour Analysis (HBA) at home for elders in AAL environments. All these applications have relevant outcomes form the current research with the ability to identify users while performing tracking, interaction analysis or HBA. Furthermore all these scenario can gather data using low cost sensors and processing units, ensuring scalability. Finally the proposed architecture can be certified on a EU basis Privacy by Design approach.

An example is the system for real-time human tracking and predefined human gestures detection using depth data acquired from Kinect sensor installed right above the detection region described in [2]. The tracking part is based on fitting an articulated human body model to obtained data using particle filter framework and specifically defined constraints which originate in physiological properties of the human body. The gesture recognition part has used the timed automaton conforming to the human body poses and regarding tolerances of the joints positions and time constraints.

For advanced analysis of human behaviours Liciotti et al. in [19] have developed a highly-integrated system. The video framework exploits vertical RGB-D sensors for people tracking, interaction analysis and users activities detection in domestic scenarios. The depth information has been used to remove the affect of the appearance variation and to evaluate users activities inside the home and in front of the fixtures. In addition, group interactions have been monitored and analysed. The audio framework has recognised voice commands by continuously monitoring the acoustic home environment.

As previously stated, another important issue to monitor and evaluate during the people tracking is the fall detection [12,15,16,22]. The solution implemented in these papers with RGB-D camera in a top-view configuration are suitable and affordable for this aim.

An automated RGB-D video analysis system that recognises human ADLs activities, related to classical daily actions is described in [21]. The main goal is to predict the probability of an analysed subject action. Thus, abnormal behaviour can be detected. The activity detection and recognition is performed using an affordable RGB-D camera. Action sequence recognition is then handled using a discriminative Hidden Markov Model (HMM).

4 Conclusion and Considerations

In this paper, our aim has been to use this state-of-the-art report to demonstrate the potential of top-view configuration for detection and tracking applications in several sub-domains, to outline key limitations and to indicate areas of technology where solutions for remaining challenges may be found. The success of RGB-D cameras can be closely linked to their affordability and for the additional rough depth information coupled with visual images that provide. These cameras have already been successfully applied in the several field to univocally identify people and to analyse behaviours and interactions. The choice of the RGB-D camera in a top view configuration is due to its greater suitability compared with a front view configuration, usually adopted for gesture recognition or even for video gaming. The top-view configuration reduces the problem of occlusions and has the advantage of being privacy preserving, because a persons face is not recorded by the camera. Starting from this, further investigation could be devoted to explore approaches more accurate and effective such as Convolutional Neural Networks or U-Net [33].

References

1. Agusta, B.A.Y., Mittrapiyanuruk, P., Kaewtrakulpong, P.: Field seeding algorithm for people counting using kinect depth image. Indian J. Sci. Technol. **9**(48) (2016)
2. Bednarık, J., Herman, D.: Human gesture recognition using top view depth data obtained from kinect sensor (2015)
3. Bevilacqua, A., Di Stefano, L., Azzari, P.: People tracking using a time-of-flight depth sensor. In: IEEE International Conference on Video and Signal Based Surveillance, AVSS 2006, pp. 89–89. IEEE (2006)
4. Bonnin, A., Borràs, R., Vitrià, J.: A cluster-based strategy for active learning of rgb-d object detectors. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1215–1220. IEEE (2011)
5. Burbano, A., Bouaziz, S., Vasiliu, M.: 3D-sensing distributed embedded system for people tracking and counting. In: 2015 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 470–475. IEEE (2015)
6. Coşkun, A., Kara, A., Parlaktuna, M., Ozkan, M., Parlaktuna, O.: People counting system by using kinect sensor. In: 2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA), pp. 1–7. IEEE (2015)

7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
8. Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., Vento, M.: Counting people by RGB or depth overhead cameras. Pattern Recogn. Lett. **81**, 41–50 (2016)
9. Dittrich, F., Woern, H., Sharma, V., Yayilgan, S.: Pixelwise object class segmentation based on synthetic data using an optimized training strategy. In: 2014 First International Conference on Networks & Soft Computing (ICNSC), pp. 388–394. IEEE (2014)
10. Felzenszwalb, P.F.: Learning models for object recognition. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I–1056. IEEE (2001)
11. Fu, H., Ma, H., Xiao, H.: Scene-adaptive accurate and fast vertical crowd counting via joint using depth and color information. Multimedia Tools Appl. **73**(1), 273 (2014)
12. Gasparrini, S., Cippitelli, E., Spinsante, S., Gambi, E.: A depth-based fall detection system using a kinect® sensor. Sensors **14**(2), 2756–2775 (2014)
13. Heath, K., Guibas, L.: Multi-person tracking from sparse 3D trajectories in a camera sensor network. In: Second ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2008, pp. 1–9. IEEE (2008)
14. Hernandez, D., Castrillon, M., Lorenzo, J.: People counting with re-identification using depth cameras (2011)
15. Kepski, M., Kwolek, B.: Detecting human falls with 3-axis accelerometer and depth sensor. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 770–773. IEEE (2014)
16. Kepski, M., Kwolek, B.: Fall detection using ceiling-mounted 3D depth camera. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 640–647. IEEE (2014)
17. Kouno, D., Shimada, K., Endo, T.: Person identification using top-view image with depth information. In: 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), pp. 140–145. IEEE (2012)
18. Liciotti, D., Contigiani, M., Frontoni, E., Mancini, A., Zingaretti, P., Placidi, V.: Shopper analytics: a customer activity recognition system using a distributed RGB-D camera network. In: Distante, C., Battiato, S., Cavallaro, A. (eds.) VAAM 2014. LNCS, vol. 8811, pp. 146–157. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12811-5_11
19. Liciotti, D., Ferroni, G., Frontoni, E., Squartini, S., Principi, E., Bonfigli, R., Zingaretti, P., Piazza, F.: Advanced integration of multimedia assistive technologies: a prospective outlook. In: 2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA), pp. 1–6. IEEE (2014)
20. Liciotti, D., Frontoni, E., Mancini, A., Zingaretti, P.: Pervasive system for consumer behaviour analysis in retail environments. In: Nasrollahi, K., Distante, C., Hua, G., Cavallaro, A., Moeslund, T.B., Battiato, S., Ji, Q. (eds.) FFER/VAAM -2016. LNCS, vol. 10165, pp. 12–23. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56687-0_2
21. Liciotti, D., Frontoni, E., Zingaretti, P., Bellotto, N., Duckett, T.: Hmm-based activity recognition with a ceiling RGB-D camera. In: Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, pp. 567–574 (2017)

22. Liciotti, D., Massi, G., Frontoni, E., Mancini, A., Zingaretti, P.: Human activity analysis for in-home fall risk assessment. In: 2015 IEEE International Conference on Communication Workshop (ICCW), pp. 284–289. IEEE (2015)
23. Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., Zingaretti, P.: Person re-identification dataset with RGB-D camera in a top-view configuration. In: Nasrollahi, K., Distante, C., Hua, G., Cavallaro, A., Moeslund, T.B., Battiato, S., Ji, Q. (eds.) FFER/VAAM -2016. LNCS, vol. 10165, pp. 1–11. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56687-0_1
24. Liciotti, D., Zingaretti, P., Placidi, V.: An automatic analysis of shoppers behaviour using a distributed RGB-D cameras system. In: 2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA), pp. 1–6. IEEE (2014)
25. Lin, S.-C., Liu, A.-S., Hsu, T.-W., Fu, L.-C.: Representative body points on top-view depth sequences for daily activity recognition. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2968–2973. IEEE (2015)
26. Liu, J., Liu, Y., Zhang, G., Zhu, P., Chen, Y.Q.: Detecting and tracking people in real time with RGB-D camera. Pattern Recogn. Lett. **53**, 16–23 (2015)
27. Lorenzo-Navarro, J., Castrillón-Santana, M., Hernández-Sosa, D.: An study on re-identification in RGB-D imagery. In: Bravo, J., Hervás, R., Rodríguez, M. (eds.) IWAAL 2012. LNCS, vol. 7657, pp. 200–207. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35395-6_28
28. Malawski, F.: Top-view people counting in public transportation using kinect. Challenges Mod. Technol. **5** (2014)
29. Marquardt, N., Hinckley, K., Greenberg, S.: Cross-device interaction via micro-mobility and f-formations. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, pp. 13–22. ACM (2012)
30. Migniot, C., Ababsa, F.: Hybrid 3D–2D human tracking in a top view. J. Real-Time Image Proc. **11**(4), 769–784 (2016)
31. Migniot, C., Ababsa, F.: 3D Human Tracking in a Top View Using Depth Information Recorded by the Xtion Pro-Live Camera. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Li, B., Porikli, F., Zordan, V., Klosowski, J., Coquillart, S., Luo, X., Chen, M., Gotz, D. (eds.) ISVC 2013. LNCS, vol. 8034, pp. 603–612. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41939-3_59
32. Rauter, M.: Reliable human detection and tracking in top-view depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 529–534 (2013)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. arXiv preprint arXiv:1505.04597 (2015)
34. Siegmund, D., Wainakh, A., Braun, A.: Verification of single-person access in a mantrap portal using RGB-D images. In: XII Workshop de Visao Computacional (WVC) (2016)
35. Tian, Q., Zhou, B., Zhao, W.-H., Wei, Y., Fei, W.-W.: Human detection using hog features of head and shoulder based on depth map. JSW **8**(9), 2223–2230 (2013)
36. Tseng, T.-E., Liu, A.-S., Hsiao, P.-H., Huang, C.-M., Fu, L.-C.: Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014), pp. 4077–4082. IEEE (2014)
37. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int. J. Comput. Vision **75**(2), 247–266 (2007)

38. Yahiaoui, T., Meurie, C., Khoudour, L., Cabestaing, F.: A people counting system based on dense and close stereovision. In: Image and Signal Processing, pp. 59–66 (2008)
39. Yamamoto, J., Inoue, K., Yoshioka, M.: Investigation of customer behavior analysis based on top-view depth camera. In: 2017 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 67–74. IEEE (2017)
40. Zhang, X., Yan, J., Feng, S., Lei, Z., Yi, D., Li, S.Z.: Water filling: unsupervised people counting via vertical kinect sensor. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 215–220. IEEE (2012)