Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration

Daniele Liciotti, Marina Paolanti(✉), Emanuele Frontoni, Adriano Mancini, 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,a.mancini,p.zingaretti}@univpm.it

Abstract. Video analytics, involves a variety of techniques to monitor, analyse, and extract meaningful information from video streams. In this light, person re-identification is an important topic in scene monitoring, human computer interaction, retail, people counting, ambient assisted living and many other computer vision research. The existing datasets are not suitable for activity monitoring and human behaviour analysis. For this reason we build a novel dataset for person re-identification that uses an RGB-D camera in a top-view configuration. This setup choice is primarily due to the reduction of occlusions and it has also the advantage of being privacy preserving, because faces are not recorded by the camera. The use of an RGB-D camera allows to extract anthropometric features for the recognition of people passing under the camera. The paper describes in details the collection and construction modalities of the dataset TVPR. This is composed by 100 people and for each video frame nine depth and colour features are computed and provided together with key descriptive statistics.

Keywords: Person re-identification · Top-view dataset · RGB-D camera · TVPR

1 Introduction

In the last decades, video analytics has been rapidly evolving as autonomous understanding of events occurring in a scene monitored by multiple video cameras. One of the fundamental problems in video surveillance is person re-identification (re-id), which is the process to determine if different instances or images of the same person, recorded in different moments, belong to the same subject. In every day life, this is done by humans without much effort. Our brains are trained to localise and detect people and later to properly re-identify them. In the recent years, this problem has gained a rapid increase in attention in both academic research communities and industrial laboratories.

Person re-id has many important applications in video surveillance, because it saves human efforts on exhaustively searching for a person from large amounts

K. Nasrollahi et al. (Eds.): VAAM 2016/FFER 2016, LNCS 10165, pp. 1–11, 2017.
DOI: 10.1007/978-3-319-56687-0_1

of video sequences. Identification cameras are widely employed in most of public places like malls, office buildings, airports, stations and museums. These cameras generally provide enhanced coverage and overlay large geospatial areas because they have non-overlapping fields-of-views. Huge amounts of video data, monitored in real time by law enforcement officers or used after the event for forensic purposes, are provided by these networks. An automated analysis of these data improves significantly the quality of monitoring, in addition to process the data faster [20].

The behaviour characterization of people in a scene and their long term activity can be possible using video analysis, which is required for high-level surveillance tasks in order to alert the security personnel.

Recent literature about re-id approaches is mostly focused on appearance-based models. Researchers have paid attention on interest points, structural information and colour as principal appearance cues [5]. The introduction of RGB-D cameras provides affordable and additional rough depth information coupled with visual images, offering sufficient accuracy and resolution for indoor applications. Due to this fact, this camera has already been successfully applied in retail field to univocally identify customers and to analyse behaviours and interactions of shoppers [12].

In this paper, we present a new dataset of person re-id that uses an RGB-D camera in a top-view configuration: the TVPR (Top View Person Re-identification) dataset. We chose an Asus Xtion Pro Live RGB-D camera because it allows acquiring colour and depth information in an affordable and fast way. The camera is installed on the ceiling above the area to be analysed.

For re-id evaluation, we collect data of 100 people, acquired across intervals of days and in different times. This choice 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 [13] and has the advantage of being privacy preserving, because the face is not recorded by the camera. Main motivations of our top-view dataset and some related applications/works are described in Table 1.

The process of extraction of a high number of significant features derived from both depth and colour information is presented. Among all possible features, we selected the nine features described in following sections as the most interesting ones. The set of features extracted by the colour and depth images is used to perform in future works the re-id process.

The paper is organized as follow: Sect. 2 is an overview of the approaches in the context of re-id; Sect. 3 gives details on the proposed setup for the collection of data, which is the core of this work; next section (Sect. 4) provides some samples and key statistics of the dataset (Subsects. 4.1 and 4.2), followed by conclusions and our future works (Sect. 5).

2 State of Art

Over the past years, in the field of object recognition a significant amount of research has been performed by comparing video sequences. Colour-based

Table 1. Main motivations and possible applications of TVPR.

Research challenges	Applications	Related works
Reliable and occlusion free people counting	Safety and security in crowded environments; people flow analysis; access control and counting	[4,11,21,24,25]
Interaction detection between people and environment	Intelligent retail environment shelf: Shopper Analytics; Ambient Assisted Living (AAL)	[6,12,16]
Fall detection, Human Behaviours Analysis (HBA)	High reliability fall detection; occlusion free; HBA at home and AAL	[10,13]

features of video sequences are usually described with the use of a set of key frames that characterize well a video sequence. The HSV colour histogram and the RGB colour histogram are robust against the perspective and the variability of resolution [9]. The clothing colour histograms taken over the head, trousers and shirt regions together with the approximated height of the person have been used as discriminative features.

Recently, the person re-id problem has received a considerable attention, and various reviews and surveys are available, pointing out different aspects of this topic [15]. Research works on person re-id can be divided into two categories: feature and learning [22].

The use of anthropometric measures for re-id was proposed for the first time in [14]. In this case, height was estimated from RGB cameras as a cue for associating tracks of individuals coming from non-overlapping views.

In [7], the authors proposed the use of local motion features to re-identify people across camera views. They obtained correspondence between body parts of different persons through space-time segmentation. On this body parts, color and edge histograms are extracted. In this approach, person re-id is performed by matching the body parts based on the features and correspondence.

Shape and appearance context, which computes the co-occurrence of shape words and visual words for person re-id is proposed in [23]. Human body is partitioned into L parts with the shape context and a learned shape dictionary. Then, these parts is further segmented into M subregions by a spatial kernel. The histogram of visual words is extracted on each subregion. Consequently, for person re-id the $L \times M$ histograms are used as visual features.

In [3] the appearance of a pedestrian is represented by combining three kinds of features (sampled according to the symmetry and asymmetry axes obtained from silhouette segmentation): the weighted color histograms, the maximally stable color regions, and recurrent highly structured patches.

Another method to face the problem of person re-id is learning discriminant models on low-level visual features. Adaboost is used to select an optimal ensemble of localized features for pedestrian recognition in [9]. Partial least squares

is used to perform person re-id in [19]. Instead, Prosser et al. [18] have used ranking SVM to learn the ranking model.

In last years, it is well-known the metric learning for person re-id. A probabilistic relative distance comparison model has been proposed [26]. It maximizes the probability that the distance between a pair of true match is smaller than that between an incorrect match pair.

In [17], the authors investigate whether the re-id accuracy of clothing appearance descriptors can be improved by fusing them with anthropometric measures extracted from depth data, using RGB-D sensors, in unconstrained settings. They also propose a dissimilarity-based framework for building and fusing the multimodal descriptors of pedestrian images for re-id tasks, as an alternative to the widely used score-level fusion.

Several datasets used to test re-id models are available: *VIPeR*[1], *iLIDS*,[2] *ETHZ*[3] and the more recent *CAVIAR4REID*[4]. These datasets cover many aspects of the person re-id problem, such as shape deformation, occlusions, illumination changes, very low resolution images, image blurring, etc. [8]. Another re-id dataset is proposed in [2]; this is composed by 79 people and four groups. Data are gathered using RGB-D technology, but are not suitable for our purposes as mentioned above in Table 1.

3 Setup and Acquisition

We have built a dataset, TVPR[5], of 100 individuals recorded from an RGB-D camera installed in a top-view configuration. The 100 people were captured in several days (see more information on TVPR in Sect. 4). The camera is installed on the ceiling of a laboratory at 4 m above the floor and covers an area of 14.66 m^2 (4.43 m × 3.31 m). The camera is positioned above the surface which as to be analysed (Fig. 1).

The first step is the processing of the data acquired from the RGB-D camera. The camera captures depth and colour images, both with dimensions of 640 × 480 pixels, at a rate up to approximately 30 fps and illuminates the scene/objects with structured light based on infrared patterns.

Seven out of the nine features selected are the *anthropometric features* extracted from the depth image:

- distance between floor and head, d_1;
- distance between floor and shoulders, d_2;
- area of head surface, d_3;
- head circumference, d_4;
- shoulders circumference, d_5;

[1] https://vision.soe.ucsc.edu.
[2] http://www.eecs.qmul.ac.uk.
[3] https://data.vision.ee.ethz.ch/cvl/aess/dataset.
[4] http://www.lorisbazzani.info/datasets.
[5] http://vrai.dii.univpm.it/re-id-dataset.

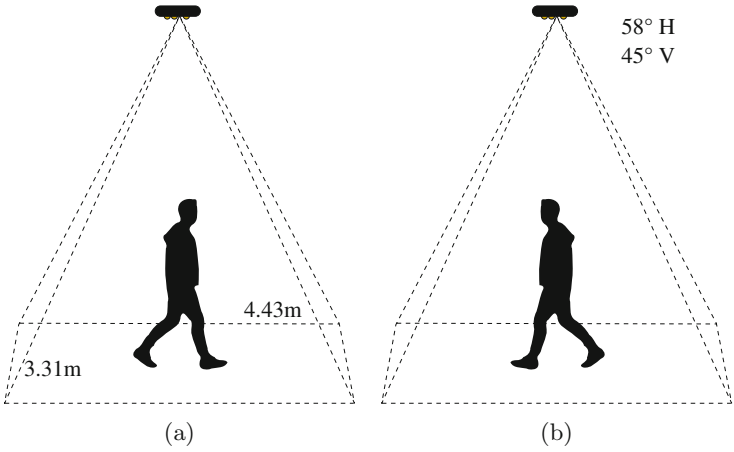

Fig. 1. System architecture.

- shoulders breadth, d_6;
- thoracic anteroposterior depth, d_7.

The remaining two *colour-based features* are acquired by the colour image. We also define *TVH*, *TVD* and *TVDH*.

- *TVH* is the colour descriptor:

$$TVH = \{H_h^p, H_o^p\} \tag{1}$$

- *TVD* is the depth descriptor:

$$TVD = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p\} \tag{2}$$

- Finally, *TVDH* is the signature of a person defined as:

$$TVDH = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p, H_h^p, H_o^p\} \tag{3}$$

Colour is an important visual attribute for both computer vision and human perception. It is one of the most widely used visual feature in image/video retrieval. To extract this two features we used HSV histograms. Local histograms have proven to be largely adopted and very effective. The signature of a person is also composed by two colour histograms computed for head/hairs and outerwear: H_h^p, H_o^p in (3), such as in [1], with $n = 10$ bin quantization, for both H channel and S channel.

Figure 2 depicts the set features considered: anthropometric and the colour-based ones.

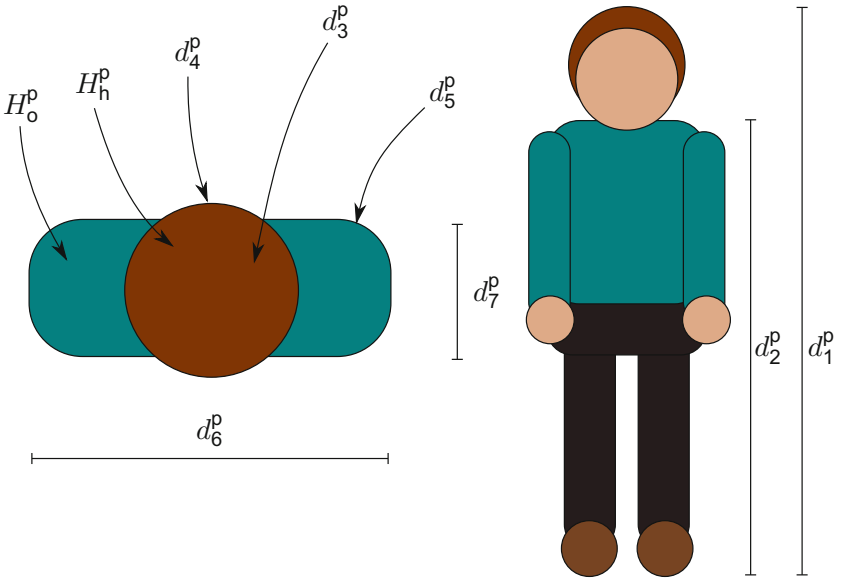

Fig. 2. Anthropometric and colour-based features.

4 Evaluation Results

4.1 Dataset Description

The 100 people of our dataset were acquired in 23 registration session. Each of the 23 folders contains the video of one registration sessions. The recording time [s] for the session and the number of persons of that session are reported in Table 2. Acquisitions have been performed in 8 days and the total recording time is about 2000 s. Registrations are made in an indoor scenario, where people pass under the camera installed on the ceiling. Another big issue is environmental illumination. In each recording session, the illumination condition is not constant, because it varies in function of the different hours of the day and it also depends on natural illumination due to weather conditions. The video acquisitions, in our scenario, are depicted in Fig. 3, which are examples of person registration respectively with sunlight and artificial light. Each person during a registration session walked with an average gait within the recording area in one direction, then it turned back and repeated the same route in the opposite direction. This methodology is used for a better split of TVPR in training set (the first passage of the person under the camera) and testing set (when the person passed again under the camera).

The recruited people are aged between 19–36 years: 43 females and 57 male; 86 with dark hair, 12 with light hair and 2 are hairless. Furthermore, of these people 55 have short hair, 43 have long hair. The subjects were recorded in their everyday clothing like T-shirts/sweatshirts/shirts, loose-fitting trousers, coats, scarves and hats. In particular, 18 subjects wore coats and 7 subjects

Table 2. Time [s] of registration for each session and the number of people of that session.

Session	Time [s]	# people	Session	Time [s]	# people
g001	68.765	4	g013	102.283	6
g002	53.253	3	g014	92.028	5
g003	50.968	2	g015	126.446	6
g004	59.551	3	g016	86.197	4
g005	75.571	4	g017	95.817	5
g006	128.827	7	g018	57.903	3
g007	125.044	6	g019	82.908	5
g008	75.972	3	g020	87.228	4
g009	94.336	4	g021	42.624	2
g010	116.861	6	g022	68.394	3
g011	101.614	5	g023	56.966	3
g012	155.338	7			
			Total	**2004.894**	**100**

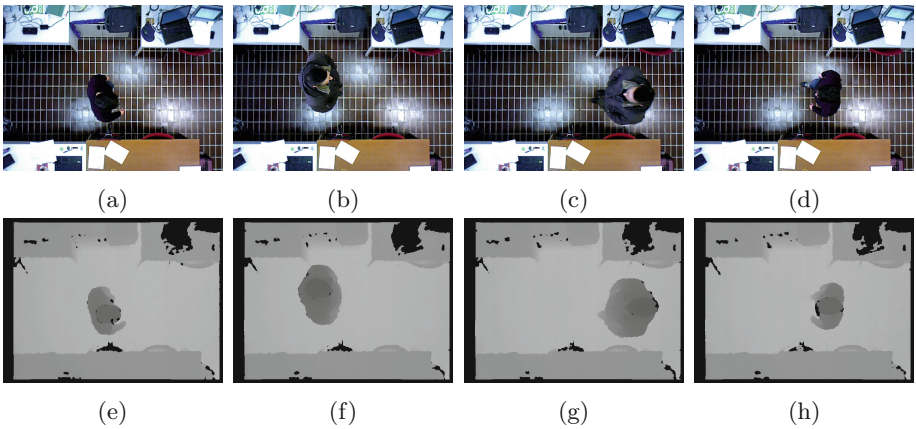

Fig. 3. Snapshots of a registration session of the recorded data, in an indoor scenario, with artificial light. People had to pass under the camera installed on the ceiling. The sequence a–e, b–f corresponds to the sequence d–h, c–g respectively training and testing set of the classes `8-9` for the registration session `g003`.

wore scarves. All videos have fixed dimensions and a frame rate of about 30 fps. Videos are saved in native `.oni` files, but can be converted in any other format. Colour stream is available in a non compressed format.

Figure 4 reports the histograms of each extracted anthropometric feature. Due to the dissimilarity of the analysed subjects a Gaussian curve is obtained from the data.

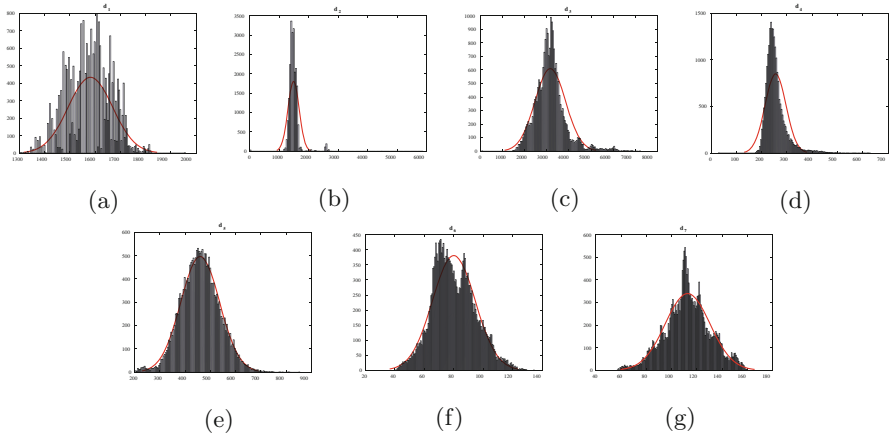

Fig. 4. Statistics histogram for each feature (**a** d_1 distance between floor and head; **b** d_2 distance between floor and shoulders; **c** d_3 area of head surface; **d** d_4 Head circumference; **e** d_5 shoulders circumference; **f** d_6 shoulders breadth; **g** d_7 thoracic anteroposterior depth). The resultant Gaussian curve (in red) is due to the dissimilarity of the analysed subjects. (Color figure online)

4.2 Performance Validation

The *Cumulative Matching Characteristic* (CMC) curve represents the expectation of finding the correct match in the top n matches. It is equivalent of the ROC curve in detection problems. This performance metric evaluates recognition problems, by some assumptions about the distribution of appearances in a camera network. It is considered the primary measure of identification performance among biometric researchers.

As well-established in recognition and in re-id tasks, for each testing item we ranked the training gallery elements using standard distance metrics. We examined the effects of 3 distance measures as the matching distance metrics: the L1 City block, the Euclidean Distance and the Cosine Distance.

To evaluate our dataset, the performance results are reported in terms of recognition rate, using the CMC curves, illustrated in Fig. 5. In particular, the horizontal axis is the rank of the matching score, the vertical axis is the probability of correct identification.

Considering our dataset, we depict a comparison among *TVH* and *TVD* in terms of CMC curves, to compare the ranks returned by using these different descriptors.

Figure 5a provides the CMC obtained for *TVH*. Figure 5b represents the CMC obtained for *TVD*. We compare these results with the average obtained by *TVH* and *TVD*. The average CMC is displayed in Fig. 5d.

It is observed that the best performance is achieved by the combination of descriptors. In Fig. 5d, it can be seen that the combination of descriptors improve the results obtained by each of the descriptor separately. This result is due to the

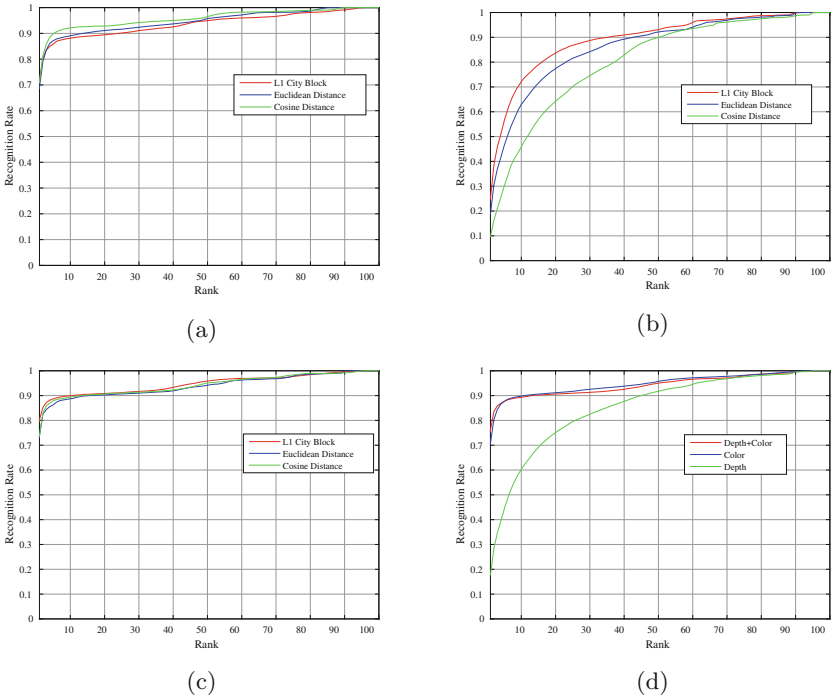

Fig. 5. The CMC curves obtained on TVPR dataset.

depth contribution that can be more informative. In fact, the depth outperform the color, giving the best performance for rank values higher than 15 (Fig. 5b). Its better performance suggests the importance and potential of this descriptor.

5 Conclusions and Future Works

Person re-identification is a critical problem in video analytics applications such as surveillance and security. In this paper, we have proposed a novel dataset for the person re-identification (TVPR) with a features set extracted from colour and depth images.

We use an RGB-D camera to detect, track and describe individuals crossing a monitored area. We chose the top-view configuration for a greater suitability, i.e. more robustness, to a series of tasks like those reported in Table 1.

Further investigation will be devoted to the study of more sophisticated features. The CMC curves have suggested that for the different distance metric approaches the depth descriptor has strong discriminative power. The integration of more features in the model seems to improve the identity discrimination. This aspect is of great importance important, in order to perform a classification model.

Future works would include the integration of this re-identification system with an audio framework and the use of other types of RGB-D sensors, such as time of flight (TOF) ones. The system can additionally be integrated as a source of high semantic level information in a networked ambient intelligence scenario, to provide cues for different problems, such as detecting abnormal speed and dimension outliers, that can alert of a possible uncontrolled circumstance.

References

1. Baltieri, D., Vezzani, R., Cucchiara, R.: Learning articulated body models for people re-identification. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 557–560. ACM (2013)
2. Barbosa, I.B., Cristani, M., Bue, A., Bazzani, L., Murino, V.: Re-identification with RGB-D sensors. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 433–442. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_43
3. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Comput. Vis. Image Underst. **117**(2), 130–144 (2013)
4. Castrillón-Santana, M., Lorenzo-Navarro, J., Hernández-Sosa, D.: People semantic description and re-identification from point cloud geometry. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 4702–4707. IEEE (2014)
5. D'Angelo, A., Dugelay, J.-L.: People re-identification in camera networks based on probabilistic color histograms. In: IS&T/SPIE Electronic Imaging, p. 78820K. International Society for Optics and Photonics (2011)
6. Frontoni, E., Mancini, A., Zingaretti, P.: RGBD sensors for human activity detection in AAL environments. In: Longhi, S., Siciliano, P., Germani, M., Monteriú, A. (eds.) Ambient Assisted Living: Italian Forum 2013, pp. 127–135. Springer, Heidelberg (2014)
7. Gheissari, N., Sebastian, T.B., Hartley, R.: Person re-identification using spatiotemporal appearance. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1528–1535. IEEE (2006)
8. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person Re-identification. Advances in Computer Vision and Pattern Recognition, 1st edn. Springer, London (2014)
9. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_21
10. 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)
11. 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)
12. 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). doi:10.1007/978-3-319-12811-5_11

13. 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)
14. Madden, C., Piccardi, M.: Height measurement as a session-based biometric for people matching across disjoint camera views. In: Image and Vision Computing New Zealand, pp. 282–286. Citeseer (2005)
15. Messelodi, S., Modena, C.M.: Boosting fisher vector based scoring functions for person re-identification. Image Vis. Comput. **44**, 44–58 (2015)
16. Migniot, C., Ababsa, F.: 3D human tracking from depth cue in a buying behavior analysis context. In: Wilson, R., Hancock, E., Bors, A., Smith, W. (eds.) CAIP 2013. LNCS, vol. 8047, pp. 482–489. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40261-6_58
17. Pala, F., Satta, R., Fumera, G., Roli, F.: Multimodal person reidentification using RGB-D cameras. IEEE Trans. Circ. Syst. Video Technol. **26**(4), 788–799 (2016)
18. Prosser, B., Zheng, W.-S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: BMVC, vol. 2, p. 6 (2010)
19. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), pp. 322–329. IEEE (2009)
20. Tu, P.H., Doretto, G., Krahnstoever, N.O., Perera, A.A., Wheeler, F.W., Liu, X., Rittscher, J., Sebastian, T.B., Yu, T., Harding, K.G.: An intelligent video framework for homeland protection. In: Defense and Security Symposium, p. 65620C. International Society for Optics and Photonics (2007)
21. Vera, P., Monjaraz, S., Salas, J.: Counting pedestrians with a zenithal arrangement of depth cameras. Mach. Vis. Appl. **27**(2), 303–315 (2016)
22. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recognit. Lett. **34**(1), 3–19 (2013)
23. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: IEEE 11th International Conference on Computer Vision (ICCV 2007), pp. 1–8. IEEE (2007)
24. Wateosot, C., Suvonvorn, N.: Top-view based people counting using mixture of depth and color information. In: The Second Asian Conference on Information Systems (ACIS) (2013)
25. 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)
26. Zheng, W.-S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: 2011 IEEE conference on Computer vision and pattern recognition (CVPR), pp. 649–656. IEEE (2011)