People semantic description and re-identification from point cloud geometry

Modesto Castrillón-Santana, Javier Lorenzo-Navarro and Daniel Hernández-Sosa
SIANI - Universidad de Las Palmas de Gran Canaria
Spain
modesto.castrillon@ulpgc.es

***Abstract*—The automatic extraction of biometric descriptors of anonymous people is a challenging scenario in camera networks. This task is typically accomplished making use of visual information. Calibrated RGBD sensors make possible the extraction of point cloud information. We present a novel approach for people semantic description and re-identification using the individual point cloud information. The proposal combines the use of simple geometric features with point cloud features based on surface normals. To test the system validity, we have collected a new and challenging dataset using a RGBD sensor in a top view configuration containing up to 63 identities captured in different sessions in different days within a two weeks period. The results achieved outperform the previous literature based exclusively on geometric features for re-identification, providing additionally very promising results in people description related to gender and hair style.**

I. Introduction

The recent arrival of affordable RGBD sensors has raised a great interest in the Computer Vision community. Current consumer depth cameras are able to provide depth information in addition to the color images. This now cheaper and compact extra information has recently been used in a number of human interaction scenarios, with remarkable results [1], including re-identification tasks [2], [3], [4], [5], [6], [7].

A possible scenario of application is related with people detection and characterization in camera/sensor networks. Automatic audience and dwell time measurement, for example, is an important issue for businesses and advertisers. Due to this relevance, many methods have been proposed to know as exactly as possible the number and characteristics of people entering/leaving a building or in public areas.

Even when the information provided by a visual sensor, e.g. 2D video, is a valuable data source to solve different computer vision problems, it also incorporates a certain degree of ambiguity that can hinder the processing. In this sense, to design a system flexible enough for any background, and observing the difficulties inherent in uncontrolled illumination for vision based background subtraction techniques, we explore the use of 3D information computed from the depth data of RGBD sensors.

Not relying in visual data permits the use of our approach in troublesome environments with highly variable and low illumination levels, such as cinemas, discos and pubs. On the contrary, stereo pair schemes face important limitations when applied to those scenarios.

Among the different possibilities, we have focused on sensors located in a zenithal configuration to simplify the problem of people detection, as occlusions are rarely present. This configuration has not been extensively used for RGBD sensors, but has exhibited its advantages in the literature using monocular and stereo systems [8], [9]. This configuration preserves privacy, and eases the detection and tracking of salient objects in scenarios subject to changing illumination conditions.

The use of a zenithal setup may restrict the field of view, but is a valid option to control access areas, for example entrances/doors, while preserving privacy and avoiding most occlusions situations. To our knowledge, there are at least two recent papers based on a zenithal RGBD sensor for re-identification purposes [6], [10]. However, we have not located any other work, using that configuration setup, related to other semantic descriptors such as gender, etc.

Therefore, this setup has proven to be adequate for people counting and more recently for re-identification. We argue that additional soft biometrics information may be extracted, being of great interest for audience analysis, not just for surveillance but also for marketing purposes.

To confirm this, an annotated database with challenging situations for re-identification in the same and different days, audience behavioral and soft biometrics analytics. We adopt the use of 3D information for that purpose, introducing the use of surface features to extract soft biometric traits, e.g. gender and hair cut, and re-identify individuals in a multi-day problem.

The main contributions of the paper are: 1) the novel approach for people semantic description and re-identification combining the use of simple geometric features with point cloud features based on surface normals, and 2) the creation of a dataset that includes depth and visual information of a set of individuals captured in 6 different sessions changing illumination and clothing attributes.

II. Scenario and Dataset

The experiments carried out in the previously mentioned works using the zenithal configuration cover a reduced number of different identities, up to 10 in [10], and up to 20 in [6]. In both cases the dataset was recorded in a single day session.

There is clearly a need of a larger dataset for this task and scenario. In this sense, we have mounted a single and stationary Kinect camera covering, from a top view, the entrance to a classroom, see Figure 1. In order to focus on

Fig. 1. Sample RGB frames extracted from the two sessions of the first day of recordings.

TABLE I. DATASET STATISTICS, OBSERVE THAT THE TOTAL NUMBER OF DIFFERENT IDENTITIES IS NOT THE SIMPLE ACCUMULATION OF THE DIFFERENT IDENTITIES IN EACH SESSION, AS DIFFERENT INDIVIDUALS ARE PRESENT IN MANY OF THEM.

Label		Session						Total
		1	2	3	4	5	6	
Crossings		31	23	34	25	21	16	312
Different ids		26	33	32	34	25	24	63*
Gender	Female	31	23	34	25	21	16	150
	Male	30	18	43	19	40	12	162
Hair length	Long	23	18	16	16	13	11	97
	Short	30	18	45	22	40	13	168
	Medium	2	2	2	0	3	1	10
	Tail	6	3	14	6	5	3	37

potentially relevant areas of the image and extract as much information as possible from significant areas in each frame, background subtraction is applied to detect salient objects.

Assuming this setup, we have recorded 2 sessions per day in 3 different days, with a one week gap, locating the camera close to a classroom entrance. The capture was launched each day before the lesson start (approx. 8 am), and after the lesson end (approx. 10 am). For each of the 6 recordings, see corresponding first frames in Figure 2, a Kinect sensor is located roughly at a similar location, approximately 2.7 meters height, looking at the scenario floor. Obviously, the illumination conditions change from one day to another, and even within the same day due to the lap between the start and the end of the class, and the sensor makes use of its auto adjustment.

The final dataset statistics for the descriptors considered in this work are summarized in Table I. Around 60 different identities are present in the whole set of recordings, however, not all of them are visible in any session. The dataset, called DEPTHVISDOOR, includes additional information of each individual related to his/her hair color, and clothing style (torso, lower body and shoes).

III. BLOB EXTRACTION

As already mentioned, and similarly to the approach presented in [6], [10], we have located the sensor in a top view configuration, with the aim at easing the segmentation. However, instead of developing a background subtraction technique based on the gray level assigned to the depth image as done by [6], we have designed a solution that works on the 3D coordinates. The approach makes use of the Point Cloud Library (PCL) [11], defining a background model over a regularized point cloud of 640×320 points.

We have adopted a simple and fast background model computed from the initial depth frames, that is robust enough for the detection and extraction problem. This background modeling solution takes advantage of the static camera configuration, and the weak influence of illumination changes in the point cloud.

To define the scene background model, bg, we make use of the mean point cloud at the session beginning. For the purpose of our experiments, this background was later not updated. The mean point cloud, $\overline{pc}$, is calculated as the average of the first k point clouds (assuming that no individual is visible and static in most frames) as:

$$bg(i,j) = \overline{pc}(i,j) = \frac{\sum_{l=1}^{k} pc^l(i,j)}{k} \tag{1}$$

where $pc^l(i,j)$ are the (x,y,x) coordinates associated to the pixel (i,j) of the $l-th$ point cloud image of the sequence.

Once the background model is available, considering that closer point cloud points have larger z, and similarly to computer graphics algorithms of surface removal, a distance test is evaluated for each pixel. Simple and fast standard background subtraction techniques may be applied. More specifically, we have adopted an approach similar to the one proposed in [12]. The foreground is computed based on a defined distance threshold applied to the background model (x,y,z) value, marking as foreground those pixels in the current point cloud whose distance with the corresponding background pixel is greater than the predefined distance threshold. More precisely, for a pixel in a given point cloud, $pc(i,j)$, its corresponding pixel in the foreground image, fg, is computed as:

$$fg(i,j) = \begin{cases} pc(i,j) & \text{if } pc(i,j) > bg(i,j) + \tau \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In our experiments, we have used $\tau = 10$ cm as distance threshold.

An example of this process is illustrated for the image presented in Figure 3. The background removal provides the image depicted in Figure 4 reporting the area where the current frame point cloud seems to be closer to the sensor. The reader may observe the presence of different detection artifacts close to those areas that may introduce spurious readings to the sensor, due to shadows. That noise is, however, simple to

Fig. 2. First RGB frame of each recording session.

Fig. 3. Input RGB image example with a single individual used to illustrate the detection and characterization process.

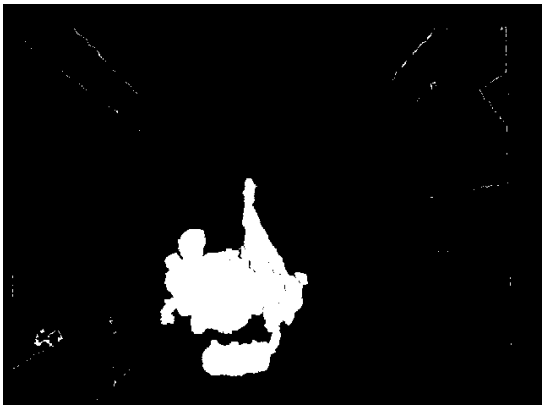

Fig. 4. Foreground mask obtained. Blob holes are basically produced to depth readings singularities.

remove as we will focus on larger blobs, see Figure 5 for the resulting sample blob mask obtained.

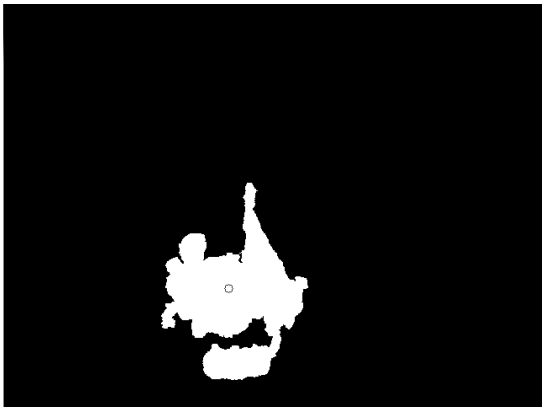

Fig. 5. Relevant blob mask.

IV. Features

A. Previous work

As argued in [10], appearance based features are sensitive to illumination changes produced, for example, by the camera auto adjustment, or the move of an individual from an indoor to an outdoor sensor location. For that reason, the authors focused on gait based features on 3D point clouds, employing a height dynamics model. Another proposal, making use of a simple approach based on 2D geometric descriptors, was described in [6]. Their descriptors are based on the salient blob height, area and volume. However, both area and volume are calculated in the pixels space, not being therefore of application if the sensor(s) is(are) located at different distances from the floor.

Recent works aiming at extracting features from point

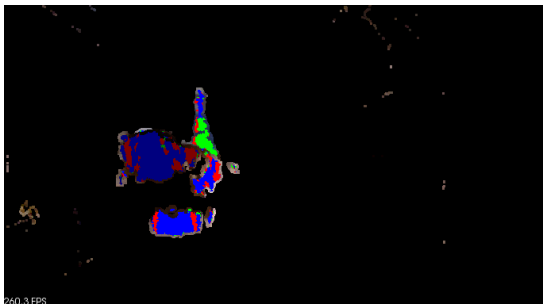

Fig. 6. Normal largest component of those points contained in the selected mask, see the mask in Figure 5, and the original RGB image in Figure 3. Respectively blue for largest z, red for largest x, and green for largest y normal component.

clouds for object recognition are making use of different descriptors. We can remark: 1) Point Feature Histograms (PFH) [13] to describe the local geometry around a point in 3D point clouds, that may introduce a too dense description; 2) SURE [14] based on interest points, analyzing the variation in surface orientation from its normals; 3) Adapting the concepts of HOG to the depth information HOD (Histogram of Oriented Depth) [15]; and 4) based on the orientation of the normal vector as HONV (Histogram of Oriented Normal Vectors) [16].

B. Our approach

In our proposal, as first step we decided to make use of the area computed in the point cloud metric space, once the sensor intrinsics are available, and therefore the calibration is performed. In this sense, the blob metric area will not be affected notoriously by the distance from the sensor to the individual. Following the ideas reported in [6] based on [17], the detected point cloud is divided into head and no-head clouds. However, the head is extracted attending to the surface geometry of the cloud instead of to a height criteria. Figure 6 presents the extracted point cloud depicting in different colors as a function of the largest normal component. The head cloud is assumed to exhibit the highest z value, and its upper part normals should be *looking at* the sensor. Therefore, to segment the head area from the point cloud, we select those points with the largest z normal component within the highest $1/7$ point cloud height range (depicted in blue in Figure 6).

Fig. 7. Simplified illustration of the surface normals for two different subjects.

Additionally, we will integrate other descriptors that are easily extracted from a point cloud, and that may exhibit a certain degree of scale invariance. Observing the human *image* perceived from the sensor, see Figure 3, in most situations there is a view of the upper torso and the head. This evidence has convinced us to evaluate the information contained in the normals present in the upper part of the body. As illustrated in Figure 7, the normals distribution will likely be different attending to the individual geometry.

The normal distribution has already been used in the literature to describe objects captured with RGBD sensors. Indeed the Histogram of Oriented Normal Vectors (HONV) [16] representation has outperformed other object descriptors such as HOG or HOG3D in object recognition tasks. This descriptor has been used for object recognition in the past, considering a grid of $n \times m$ cells that produce a concatenated histogram for object description.

Certainly, this representation offers different advantages. For that reason we propose to extract features from a selected area of the point cloud, making use of a histogram based representation. However, we will not consider the grid of histograms, as the individual, and his head head, seen from above, must be represented in such a way that the invariance to 2D rotations is preserved. Additionally, we have adopted a representation based exclusively on the zenith angle employed in [16], i.e. the angle of the normal with respect to the xy plane. We have limited the histogram to 18 bins without any significant loss in accuracy, while reducing the problem dimensionality.

To summarize, we will consider different basic features:

- Height (H)
- Area in pixels (A)
- Volume in $pixels \times meters$ (V)
- Area in metric space (A_m)
- Histogram of zenith angles ($Hist$)

As indicated below with more details in the experiments section, the histogram of zenith angles is obtained for a specific range of the blob height.

The following feature combinations have been analysed in the experimental setup:

- Height (H)
- Height and histogram of zenith angles ($HHist$)
- Height and area in pixels (HA)
- Height, area in pixels, and volume in $pixels \times meters$ (HAV)
- Height, area in pixels, volume in $pixels \times meters$, and histogram of zenith angles ($HAVHist$)
- Height and area in metric space (HA_m)
- Height,area in metric space and histogram of zenith angles (HA_mHist)

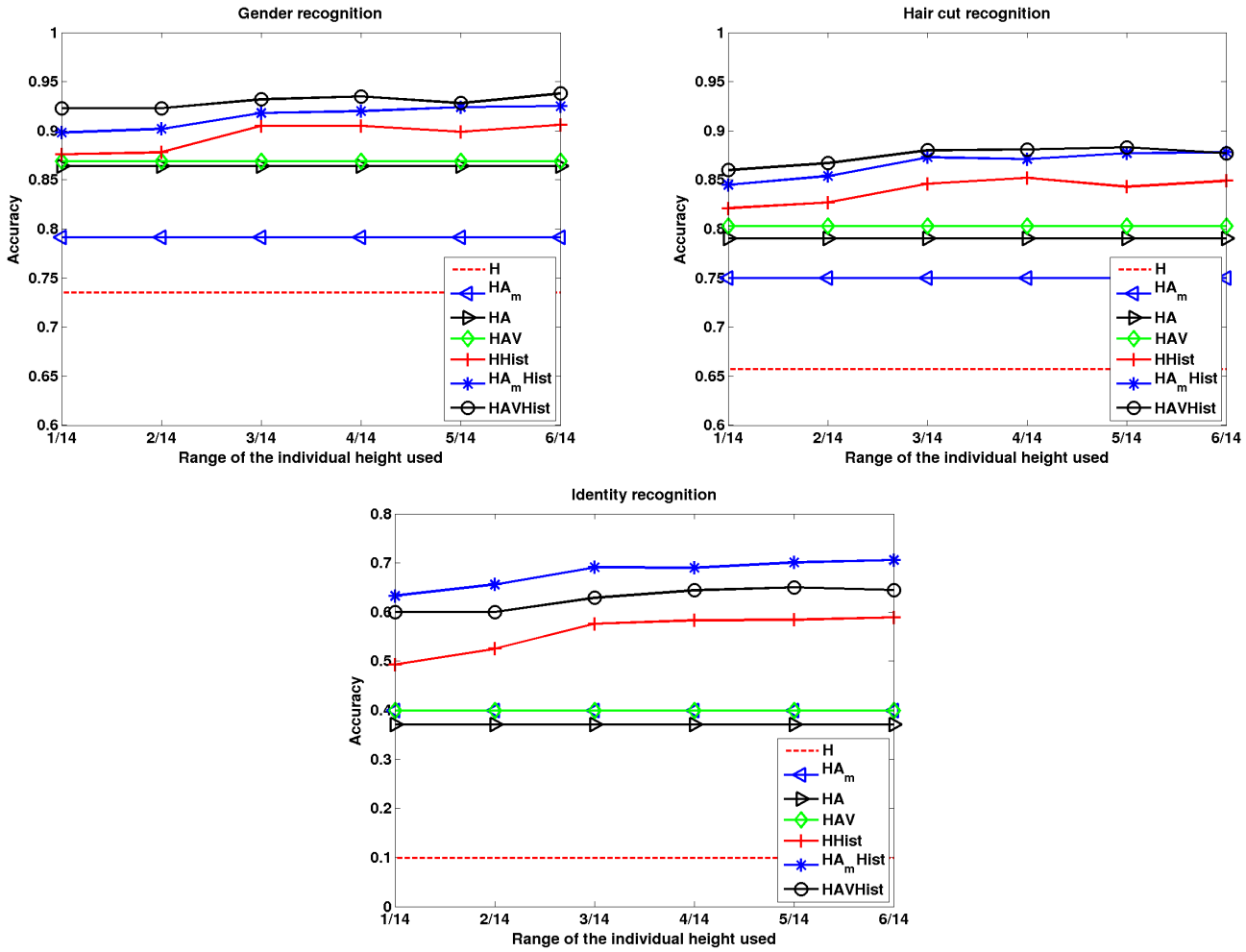

Fig. 8. Results achieved for the different feature combinations in the defined problems. Observe that the height range modifies the performance only for those features vectors that make use of the zenith angles histogram.

V. Experimental Setup and Results

In the experiments, we have investigated three different biometric traits: gender, hair cut and identity. The first two are certainly soft biometric traits, with not immediate application for re-identification tasks, but of particular interest for audience analysis applications. For this task, we have analyzed in a 5-folds setup the different sets of features. According to this, the experiment is repeated 5 times, testing each time with a different fold, and training with the remaining 4 folds.

For those experiments involving the use of the normal zenith angles, the corresponding histogram is computed considering not the whole point cloud of an individual, but a specific height range starting from the point closest to the camera (likely the head), i.e. the point with the largest z component in the cloud. As denoted by Leonardo da Vinci in his Vitruvian Man, based on the work of Vitruvius, in the ideal human proportions *from below the chin to the top of the head is one-eighth of the height of a man*. This observation will inspire the method used to determine the point cloud region of interest for identity, gender and hair cut recognition in the experiments section.

In the experiments we have evaluated the performance achieved for 6 different ranges of the individual height. Beginning with a initial height range of $1/14$ of the individual cloud height, i.e. the upper head part, and with a step of $1/14$ up to a range of $6/14$ of the height in the last test, likely reaching the torso area. This height range variation has not been used in those feature combinations that do not consider the normal zenith angles.

To carry out the experiments, we have used Weka [18], testing different classification approaches, reporting here the results achieved for Random Forest, as they provided the best accuracy during our evaluation

For gender and hair cut classification, we have performed the experiment with the whole dataset, i.e. using the 6 different sessions, see Table I. The first row of Figure 8 presents the performance achieved with different feature combinations for both classification problems. The integration of the zenith angle information improves the classification performance, that is even more evident if combined with the area information. Both $HAVHist$ and HA_mHist, reach similar performance, but the reader must observe that the use of the area and volume from the pixels information ($HAVHist$) is highly affected by the scale depending on the distance from the sensor to the

floor. The advantage of using the metric area is evident as the representation will be more invariant to scale, however, the results are rather similar for all the traits, as in these experiments, the sensor is located in quite similar position in the different 6 sessions.

More precisely, for gender classification the accuracy is roughly 93% for the dataset with similar results for both classes. In the case of hair cut classification, the resultant value is almost 88%, but those classes with less samples (medium and tail) present lower accuracy.

Related with identity recognition, we have first removed from the dataset those identities being present more than 10 times (too much) or just one (too few). The final number of identities analyzed is 41. Comparing with the results achieved with HAV, ours outperform them, jumping from 40% to over 70%. The benefits of the zenith angle histogram are much more evident than in the other classification problems.

In summary, the introduction of the histogram of zenith angles improves performance results based on simple geometric features, that have already been used in the literature for identity recognition. The behavior is also noticeable for gender and hair cut classification, but for identity recognition the improvement is even more relevant. These results are achieved evaluating the gathered dataset that presents unrestricted illumination conditions. Indeed, the single use of the histogram of zenith angles is enough to beat the approach based on geometric features. Even more, the integrated information is complementary as the fusion of geometry and the zenith angle information reports even better accuracies.

Additionally the extraction of features from a larger point cloud range, seems to add information, however, the reader must observe that the visible points for a zenithal setup are each time less significant due to self-occlusion of the point cloud.

VI. Conclusions

In this paper, we propose a novel approach for people semantic description and re-identification. The system is designed to fuse simple geometric features, that have been previously tested in re-identification, with point cloud features.

The new and challenging DEPTHVISDOOR dataset[1] has been built in order to evaluate the validity of the proposal. The annotated dataset have been acquired in three different days, making a total of 6 sessions, including behavioral, identity and soft biometrics information. DEPTHVISDOOR has been registered with a RGBD sensor in a top-view configuration.

Illumination changes are frequent in the dataset, making standard color and appearance based features unsuitable for re-identification and soft biometric recognition purposes. Thus, we have compared the use of point cloud features based on the surface normals with simple geometric features proposed in previous works. The proposed approach outperforms the results obtained by geometric features based solutions. The experiments reported high accuracies in the gender and hair cut classification problems, improving significantly the performance in the problem of re-identification, reaching an accuracy of 94%, 88% and 72% respectively.

[1]The dataset may be found at http://berlioz.dis.ulpgc.es/roc-siani/descargas

As a first conclusion, these results indicate the validity of fusing simple geometric and point cloud features applied to the people description and recognition problem. However, the main conclusion suggests that geometric and point cloud features add complementary and useful information to appearance based approaches in different scenarios.

Acknowledgment

Work partially funded by the Institute of Intelligent Systems and Numerical Applications in Engineering and the Computer Science Department at ULPGC.

References

[1] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, October 2013.

[2] A. Albiol, A. Albiol, J. Oliver, and J. Mossi, "Who is who at different cameras: people re-identification using depth cameras," *IET Computer Vision*, vol. 6, no. 5, pp. 378–387, September 2012.

[3] B. I. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *1st International Workshop on Re-Identification*, 2012.

[4] J. Oliver, A. Albiol, and A. Albiol, "3D descriptor for people re-identification," in *21st International Conference on Pattern Recognition (ICPR)*, 2012.

[5] R. Satta, F. Pala, G. Fumera, and F. Roli, "Real-time appearance-based person re-identification over multiple kinect cameras," in *8th International Conference on Computer Vision Theory and Applications (VISAPP)*, Barcelona, Spain, 2013.

[6] J. Lorenzo-Navarro, M. Castrillón-Santana, and D. Hernández-Sosa, "On the use of simple geometric descriptors provided by RGB-D sensors for re-identification," *Sensors*, vol. 13, no. 7, pp. 8222–8238, 2013.

[7] J. Oliver, A. Albiol, A. Albiol, and J. Mossi, "Re-identifying people in the wild," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[8] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Computer Vision and Pattern Recognition*, 2008, pp. 1 – 7.

[9] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Transactions in Image Processing*, vol. 21, no. 4, pp. 2160–2177, April 2012.

[10] V. John, G. Englebienne, and B. Krose, "Person re-identification using height-based gait in colour depth camera," in *IEEE International Conference on Image Processing*, 2013, pp. 3345–3349.

[11] Open Perception, "Point cloud library (pcl)," http://pointclouds.org/.

[12] J. Heikkila and O. Silven, "A real-time system for monitoring of cyclists and pedestrians," in *IEEE Workshop on Visual Surveillance*, Fort Collins, Colorado, June 1999, pp. 82–90.

[13] R. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *IEEE International Conference on Robotics and Automation*, 2009.

[14] T. Fiolka, J. Stueckler, D. A. Klein, D. Schulz, and S. Behnke, "SURE: Surface entropy for distinctive 3D features," in *Spatial Cognition*, 2012.

[15] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2011.

[16] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *11th Asian Conference on Computer Vision*, 2012.

[17] G. Englebienne and B. Krose, "Fast bayesian people detection," in *22nd Benelux Conference on Artificial intelligence*, 2010.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.