Gait Recognition with Compact Lidar Sensors

Bence Gálai[1], Csaba Benedek[12]
[1]*Institute for Computer Science and Control, H-1111, Budapest, Kende u. 13-17, Hungary*
lastname.firstname@sztaki.mta.hu

[2] *Péter Pázmány Catholic University, H-1083, Budapest, Práter utca 50/A, Hungary*

Keywords: Gait Recognition, Lidar

Abstract: In this paper, we present a comparative study on gait and activity analysis using LiDAR scanners with different resolution. Previous studies showed that gait recognition methods based on the point clouds of a Velodyne HDL-64E Rotating Multi-Beam LiDAR can be used for people re-identification in outdoor surveillance scenarios. However, the high cost and the weight of that sensor means a bottleneck for its wide application in surveillance systems. The contribution of this paper is to show that the proposed Lidar-based Gait Energy Image descriptor can be efficiently adopted to the measurements of the compact and significantly cheaper Velodyne VLP-16 LiDAR scanner, which produces point clouds with a nearly four times lower vertical resolution than HDL-64. On the other hand, due to the sparsity of the data, the VLP-16 sensor proves to be less efficient for the purpose of activity recognition, if the events are mainly characterized by fine hand movements. The evaluation is performed on five tests scenarios with multiple walking pedestrians, which have been recorded by both sensors in parallel.

1 INTRODUCTION

A study in the 1960s (Murray, 1967) showed that people can recognize each other by the way they walk. Since then gait as a biometric feature has been extensively studied. Gait analysis may not be as much accurate as fingerprint or iris recognition for people identification, yet it has some benefits versus other biometric modalities. In particularly, gait can be observed from a distance, and people do not need to interact with any devices, they can just walk naturally in the field of interest. Since a single imaging sensor is enough for recording gait cycles, gait analysis can easily be adopted to surveillance systems.

Challenges with optical camera based gait recognition methods may arise from various factors, such as background motion, illumination issues and view-dependency of the extracted features. Although view-invariant (3D) descriptors can be obtained from multi-camera systems, the installation and calibration of such systems may be difficult for ad-hoc events. We can find several approaches in the literature relying on optical cameras, however their efficiency is usually evaluated in controlled test environments with limited background noise or occlusions effects. The number of practical applications where the circumstances satisfy these constraints is limited. In realistic surveillance scenarios we must expect multiple people walking with intersecting trajectories in front of a dynamic background. We need therefore view-invariant, occlusion-resistant robust features which can be evaluated in real time enabling immediate system response.

A Rotating Multi-Beam (RMB) LiDAR sensor can provide instant 3D data from a field-of-view of 360° with hundreds of thousands of points in each second. In such point clouds view invariance can be simulated with proper 3D transformations of the point cloud of each person (Benedek et al., 2016), while occlusion handling, background segmentation and people tracking can also be more efficiently implemented in the range image domain, than with optical images. (Benedek, 2014) showed that a 64-beam LiDAR (Velodne HDL-64E) is able track several people in realistic outdoor surveillance scenarios, and (Benedek et al., 2016) showed that the same sensor is also effective in the re-identification of people leaving and re-entering the field-of-view. However, the 64-beam sensor is too heavy and expensive for wide usage in surveillance systems. In this paper, we demonstrate that even lower resolution, thus cheaper LiDAR sensors are capable of accurate people tracking and re-identification, which fact could benefit the security sector, opening doors for the usage of LiDARs in fu-

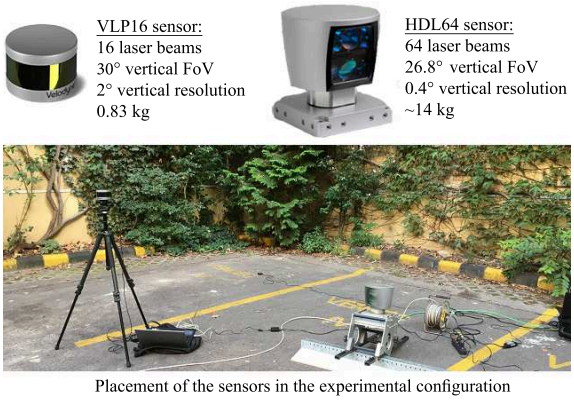


Figure 1: Main features of the used RMB LiDARs, and positioning of the sensors in the experiments.

ture surveillance systems.

The rest of the paper is organized as follows: Section 2. provides some information about related work in the field of gait recognition, Section 3. presents a brief introduction to our gait recognition method using Rotating Multi-Beam LiDAR sensor. Section 4. gives quantitative results about the accuracy of each sensor in the different gait sequences. In Section 5. experiments on activity recognition are presented. Conclusion is provided in Section 6.

2 RELATED WORK

Gait recognition has been extensively studied in the recent years (Zhang et al., 2011). The proposed methods can be divided into two categories: *model based* methods, which fit models to the body parts and extracts features and parameters like joint angles and body segment lengths, and *model free* methods, where features are extracted from the body as a whole object. Due to the characteristics and the density of point clouds generated by a Rotating Multi-Beam LiDAR sensor, like the Velodyne HDL-64E or the VLP-16, robust generation of detailed silhouettes are hard to accomplish, so we decided to follow a *model free* approach as the *model based* methods need precise information on the shape of body parts, such as head, torso, tigh etc. as described in (Yam and Nixon, 2009), which are often missing in RMB LiDAR-based environments.

There are many gait recognition approaches published in the literature which are based on point clouds (Tang et al., 2014; Gabel et al., 2012; Whytock et al., 2014; Hofmann et al., 2012), yet they use the widely adopted Kinect sensor which has limited range and a small field-of-view and is less efficient for applications in real life outdoor scenarios than LiDAR sensors. Also the Kinect provides magnitudes higher density than an RMB LiDAR, so the effectiveness of these approaches are questionable in our case.

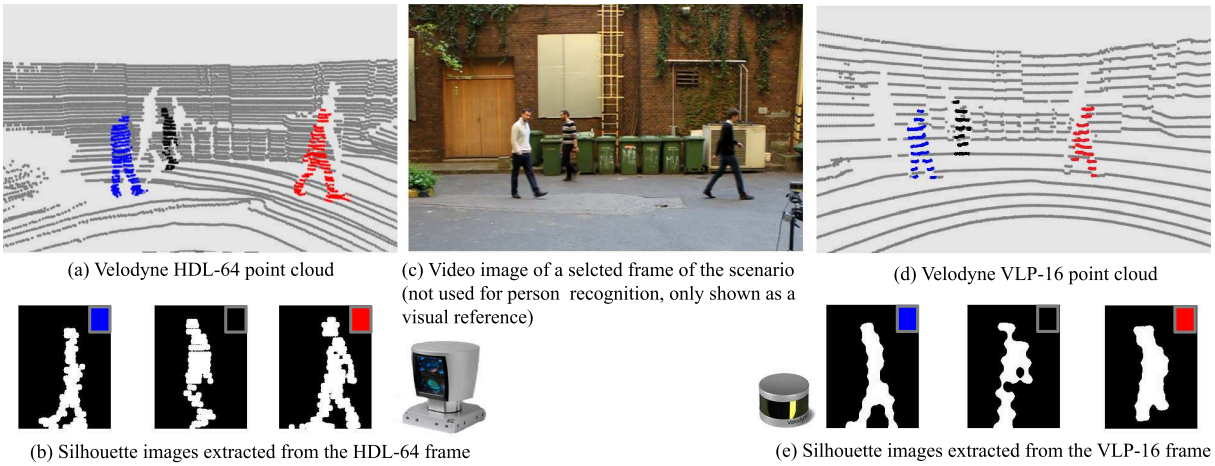
The Gait Energy Image (Han and Bhanu, 2006), originally proposed for optical video sequences, is often used in its basic (Shiraga et al., 2016) or improved version (Hofmann et al., 2012), since it provides a robust feature for gait recognition. In (Gálai and Benedek, 2015) many state-of-the-art image based descriptors were tested for RMB LiDAR point cloud streams, proposed methods for both optical images (Kale et al., 2003) and point clouds were evaluated. (Tang et al., 2014) uses Kinect point clouds and calculates 2.5D gait features: Gaussian curvature, mean curvature and local point density which are combined into 3-channel feature image, and uses Cosine Transform and 2D PCA for dimension reduction, but this feature needs dense point clouds for curvature calculation, thus not applicable for RMB LiDAR clouds. (Hofmann et al., 2012) adopts the image aggregation idea behind the Gait Energy Image and averages the pre-calculated depth gradients of a depth image created from the Kinect points. This method proved to be more robust for sparser point clouds, yet it was outperformed by the Lidar-based Gait Energy Image, which is described in Section 3. in detail.

2.1 Gait Databases

The efficiency of the previously proposed methods are usually tested on public gait databases like the CMU Mobo (Gross and Shi, 2001), the CASIA (Zheng et al., 2011) or the TUM-GAID (Hofmann et al., 2014) database. However these datasets were recorded with only a single person present at a time, with limited background motion and illumination issues, which constraints are often not fulfilled in realistic outdoor scenarios. To overcome the domination of such databases (Benedek et al., 2016) published the SZTAKI-LGA-DB dataset recorded with RMB LiDAR sensor in outdoor environments. During the experiments presented in Section 4 we followed the same approach by recording the point cloud sequenes.

2.2 Devices Used in Our Experiments

The LiDAR devices used here are the Velodyne HDL-64E and VLP-16 sensors, shown in Fig. 1. The HDL-64E sensor has a vertical field-of-view of 26.8 with 64 equally spaced angular subdivisions, and approximately 120 metres range providing more than two million points per second. The VLP-16 has 30 vertical field-of-view, 2 vertical resolution and a range of



(a) Velodyne HDL-64 point cloud

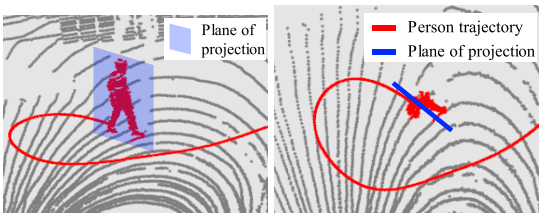
(c) Video image of a selcted frame of the scenario (not used for person recognition, only shown as a visual reference)

(d) Velodyne VLP-16 point cloud

(b) Silhouette images extracted from the HDL-64 frame

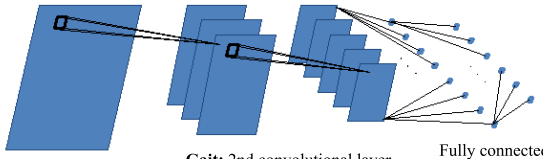
(e) Silhouette images extracted from the VLP-16 frame

Figure 2: Point clouds captured with the HDL-64E (left) and VLP-16 (right) and the associated side-view silhouettes of the three people present in the scene.



(a) (b)

Figure 3: The projection plane for LGEI generation from a) side-view, b) top view.



Parameters	k1	f1	k2	f2	h	n
gait recognition	3	5	7	9	98	*N*
activity recongition	7	5	2	-	20	1

Figure 4: Structure of the used convolutional neural networks (CNN). By gait recognition, *N* is equal to the number of people in the training set.

100 metres. Both sensors have a rotational rate of 5 Hz - 20 Hz. During the experiments, the sensors were positioned close to each other (Fig. 1, bottom), which could capture the scenario in parallel with two similar viewpoints (Fig. 2).

3 PROPOSED GAIT RECOGNITON APPROACH

In this section we present a brief introduction to the adopted gait recognition method, called the Lidar-based Gait Energy Image (LGEI).

LGEI proved to be the most effective feature for LiDAR-based gait recognition in (Gálai and Benedek, 2015). The LGEI adopts the idea of the Gait Energy Image (Han and Bhanu, 2006), by averaging side-view silhouettes in a full gait cycle, with some small yet significant alternations.

First, an LGEI is generated by averaging 60 consecutive silhouettes, which is equivalent to nearly 3-4 gait cycles, as the frame rates of the considered RMB LiDAR sensors are lower than in cases of optical cameras.

Second, since occlusions occur in the realistic outdoor scenarios of the experiments, each frame where only partial silhouettes were visible are discarded. This filtering step results in a drop of 10-12% of the training and testing images, yet it can boost the performance of the correct re-identifications.

Third for classification, the LGEI approach uses the committee of a convolutional neural network (CNN) and a multilayer perceptron (MLP). Although the neural networks require in general large amounts of input data, the convolutional network we designed was small enough, so that it could learn efficient biometric features based on a few thousand of input LGEIs withing the test set. For the multilayer perceptron, the input data was preprocessed similarly to the approach in (Han and Bhanu, 2006): principal component analysis and multiple discriminant analysis were applied to the LGEIs to create the input for the MLP. Both the CNN and the MLP used downscaled image maps of 20×15 pixels and both net-

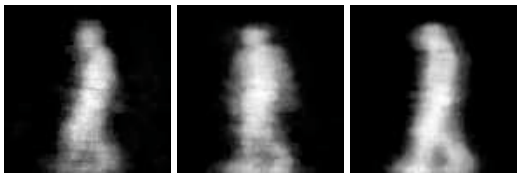
works have an output layer of N neurons, which is equal to the number of people present in the scene.

We used *tanh* activation function whose output is in the [-1,1] domain, thus for the ith person in a test scenario the network's output should be 1 for the associated neuron and -1 for all others. In the recognition phase the trained networks produce output vectors o_{cnn} and $o_{mlp} \in R^N$ in the [-1,1] domain, for the output of the CNN-MPP committee we then take the vector $o = max(o_{cnn}, o_{mlp})$. For a given G probe LGEI we then calculate $i_{max} = argmax_i(o)$ and sample G is recognized as person i_{max}, if $o^{i_{max}} > 0$, otherwise we mark G as unrecognized.

The structure of our convolutional network can be seen in Figure 4. We note here that (Wolf et al., 2016) also uses CNN for gait analysis, and the authors of (Shiraga et al., 2016) use CNNs with the Gait Energy Image inputs for classification. However, the structures of their networks is larger than the one presented here, and they also rely on a much larger dataset of optical image data (Makihara et al., 2012) for GEI generation and training.

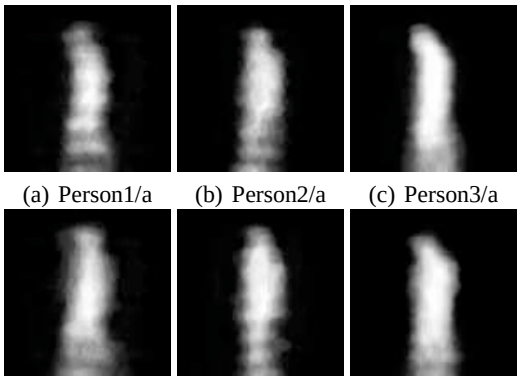
For LGEI generation the point clouds of each person are projected to a plane tangential to the person's trajectory (see Fig. 3) and morphological operations are applied to obtain connected silhouettes. Naturally in the VLP-16 sequences, even more steps of morphological post processing operations are needed to obtain connected silhouette blobs, thus in terms of level of details, the quality of the VLP-16 feature maps are notably lower than experienced with the HDL-64E point clouds. Three silhouettes extracted from a sample frame are shown in Figure 2 for visual comparison. In both the HDL-64 and VLP-16 cases, the projected silhouette images are upscaled to 200×150 pixels. In the post processing phase, the HDL-64E feature map undergoes a single dilation step with a kernel of 5×5 pixels. The same kernel is used initially for the VLP-16 silhouettes, which is followed by five cycles of alternately applying dilation and erosion kernels with a size of 3×5.

We can visually compare the LGEIs extracted from the HDL-64 and VLP-16 sequences in Fig. 5 and 6. Most important differences can be observed in the arm and leg regions, where the low-resolution sensor can only preserve less details. On the other hand, the main silhouette shape and the characteristic posture still remains recognizable even on the VLP-16 measurement maps, which fact can be confirmed by comparing different LGEIs of the same subjects in Fig. 6.



(a) Person1 (b) Person2 (c) Person3

Figure 5: HDL64-LGEI sample images



(a) Person1/a (b) Person2/a (c) Person3/a

(d) Person1/b (e) Person2/b (f) Person3/b

Figure 6: VLP16-LGEI samples: images in the same column correspond to the same person

4 EXPERIMENTS ON GAIT RECOGNITION

Out tests set consists of five scenarios containing multiple pedestrians walking in a courtyard. Each scenario was recorded by both the HDL-64 and VLP-16 sensors in parallel (see Fig. 2). In the sequences N3/1, N3/2, and N3/3 the same three test subjects were walking in the field of view with intersecting trajectories, and the VLP-16 sensor has been placed a several metres closer to the walking area than the HDL-64. Sequences F4 and F5 represent similar scenarios with four and five people, respectively, but the two devices were placed here in approximately equal, and relatively far distances from the moving people. A snapshot from the sensor configuration capturing the F4 and F5 sequences is shown in Fig. 1.

Similarly to (Benedek et al., 2016), we divided the captured sequences into distinct parts, for training and test purposes, respectively. In the near-to-sensor setting scenario (N3) the three sections are evaluated with cross validation, e.g. by testing the recognition on the N3/2 part, the training set was generated from the N3/1 segment (corresponding result is shown in Table 1, 1st row) and so on. On the other hand, the F4 and F5 sequences were split into two parts, and in both cases, the first segments were used for training and the second ones for testing the recognition per-

Table 1: Rates of correct re-identifications with the HDL-64E and VLP-16 sensors in five sequences. The scenarios N3/1, N3/2 and N3/3 were recorder while three people were walking near to the sensor, F4 and F5 with four and five people respectively far from the sensor.

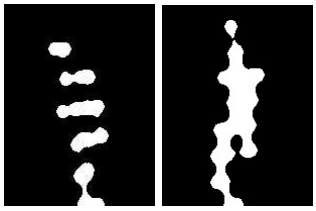
Sequence	HDL-64	VLP-16
N3/1	96%	81%
N3/2	85%	84%
N3/3	93%	81%
F4	79%	68%
F5	93%	54%

formance.

For the gallery set generation, $k = 100$ random key frames were selected from the training sequences, and the training LGEIs were calculated from the $l = 60$ consecutive silhouette images. As for the probe set, 200 seed frames were selected from the test set, and each of the 200 test LGEIs were matched independently to the trained models.

For each test scenario, the accuracy rates of correct re-identification with both sensors are shown in Table 1. As expected, the tests with HDL-64 data outperform the VLP-16 cases due to the 4-times larger vertical resolution of the point clouds, however in the near-to-sensor configuration (N3 sequences), the performance of the compact VLP-16 LiDAR can still be regarded as quite efficient (above 80%). On the other hand, for the far-from-sensor (F4 and F5) cases, the tests with the VLP-16 sensor yielded notably lower scores, which observation is the consequence of the poor measurement density from the subjects at larger distances. To demonstrate the differences between data of the two sensor configurations, we show in Fig. 7 two *worst case* silhouette examples from the far and near scenarios, respectively. While in the near-to-sensor example, the shape of the extracted human body is strongly distorted, the silhouette blob is at least still connected. On the other hand. in the far-from-sensor sequences there are many silhouette candidates, which cannot be connected even by applying several morphological operations, and consist of disconnected floating blobs. We can conclude from these experiences, that the VLP-16 sensor can indeed be applicable in future surveillance systems, however the appropriate positioning of the sensor is a key issue, as the performance quickly depreciates by increasing the distance[1].

[1]Demo videos of person tracking with various Velodyne sensors can be found in our website: http://web.eee.sztaki.hu/i4d/demo_surveillance_persontracking.html



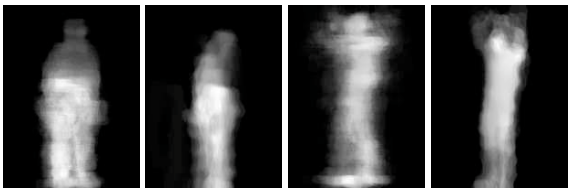
(a) (b)

Figure 7: Worst-case VLP-16 silhouettes in: a) far, b) near sensor setting recordings.

5 EXPERIMENTS ON ACTIVITY RECOGNITION

Apart from person identification, the recognition of various events can provide valuable information in surveillance systems. For activity recognition the averaging idea of Gait Energy Image can also be adopted: (Benedek et al., 2016) introduced two feature images: the Averaged Depth Maps (ADM), and the Averaged eXcluse-OR (AXOR) images. Each feature image was generated based on 40 consecutive LiDAR frames (from sequences with 10fps), which was the average duration of the activities of interest. We used frontal silhouette projections in this case, since activities were better observed from a frontal point of view. Apart from normal walk, we have selected five events for recognition: bend, check watch, phone call, wave and wave two-handed (wave2) actions.

Recording the motion of limbs in 3D is essential in the recognition of the above typical events. Since binarized silhouettes do not provide enough details for automatic analysis, depth maps were derived from the point clouds for capturing the appearance of the body. The ADM feature has been obtained by averaging the consecutive depth maps during the action, similarly to GEI calculation. An activity can also be described from it's dynamics, highlighting the parts where the frontal *depth silhouettes* change significantly in time. Thus we have derived a second feature map, so that for each consecutive frontal silhouette pairs the exclusive-OR (XOR) operator was applied capturing the changes in the contour, and by averaging the consecutive XOR images we derived the AXOR map. For recognition two convolutional neural network were used, one for the ADM and one for the AXOR image. Recording the motion of limbs in 3D is essential in the recognition of the above typical events. Since binarized silhouettes do not provide enough details for automatic analysis, depth maps were derived from the point clouds for capturing the appearance of the body. The ADM feature has been obtained by averaging the consecutive depth maps during the action, similarly to GEI



(a) HDL-64 (b) VLP-16 (c) HDL-64 (d) VLP-16

Figure 8: Good quality Averaged Depth Maps (ADM) for bend (a-b) and wave2 (c-d) actions with the two Lidar sensors.

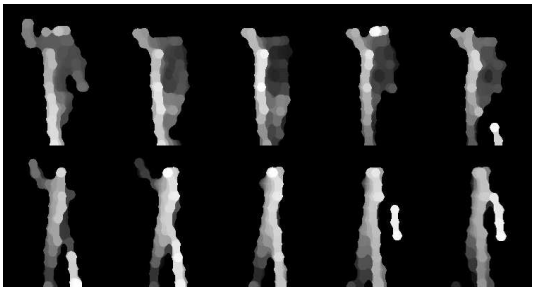
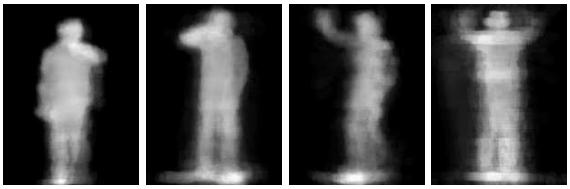


Figure 9: 10 consecutive frames of a waving activity recorded by the VLP-16 sensor.

calculation. An activity can also be described from it's dynamics, highlighting the parts where the frontal *depth silhouettes* change significantly in time. Thus we have derived a second feature map, so that for each consecutive frontal silhouette pairs the exclusive-OR (XOR) operator was applied capturing the changes in the contour, and by averaging the consecutive XOR images we derived the AXOR map. For recognition two convolutional neural network were used, one for the ADM and one for the AXOR image.

We have performed the activity recognition experiments with both LiDARs in the near-to-sensor configuration. Fig 8 shows two ADM examples – one for the *bending* and one for the two handed waving (*wave2*) action – where the qualities of the VLP-16 feature maps are similar to the HDL-64 cases. In general, the bending action could be efficiently detected by the VLP-16 sensor, but the remaining activities often struggled with the issues of low resolution. Figure 9. highlights this phenomenon: 10 consecutive frames of a waving activity are shown. We can see that the waving hand randomly disappears and reappears throughout the frames, thus in the averaging step it may be canceled out without causing characteristic patterns in the ADM and AXOR images. In Figures 10. and 11. we can see ADMs of four activities of interest derived from the measurements of the HDL-64 and the VLP-16 sensors respectively. The loss of important details between each pair of corresponding HDL-64 and VLP-16 sample images is visible in the figures, these VLP-16 ADMs are difficult



(a) watch (b) phone (c) wave (d) wave2

Figure 10: Reference ADMs generated from HDL-64E clouds

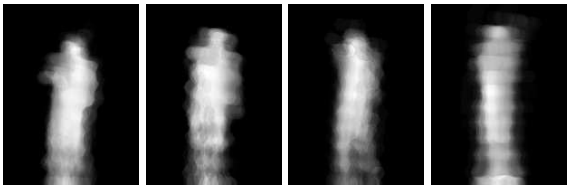
to distinguish even by human observers. While the measured recognition rates were above 85% in the HDL-64 sequences (Benedek et al., 2016), we have concluded hereby that for reliable recognition of precise hand movements in the ADM/AXOR feature image domain, the 2° vertical resolution of the basic VLP-16 sensor is less efficient. However, as the tendency in the compact sensor development indicates the increase of the vertical resolution parameter versus the field of view (the newest model of the company reaches 1.33° resolution within a 20° FoV), the doors for this particular application will be soon open for compact Lidar sensors as well.

6 CONCLUSION

We showed that the gait recognition task can be efficiently approached with low resolution RMB LiDARs like the VLP-16 sensor. The proposed gait recognition method was able to achieve a relatively high accuracy, since it uses the motion of the whole body as descriptor. We also showed that the distance of the VLP-16 sensor from the walking people largely influence the results, but with precise positioning of the device could accomplish similar performance to ones acquired from the HDL-64. On the other hand, various activity recognition functions based on principally hand movements face limitations by the low density VLP-16 point clouds, and we experienced larger gaps in recognition performance between the two sensors. This work was supported by the National Research, Development and Innovation Fund (NKFIA #K-120233). C. Benedek also acknowledges the support of the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

REFERENCES

Benedek, C. (2014). 3D people surveillance on range data sequences of a rotating Lidar. *Pattern Recognition Letters*, 50:149–158. Special Issue on Depth Image Analysis.



(a) watch (b) phone (c) wave (d) wave2

Figure 11: Low quality ADM samples generated from VLP-16 clouds for the actions of Fig. 10

Benedek, C., Gálai, B., Nagy, B., and Jankó, Z. (2016). Lidar-based gait analysis and activity recognition in a 4D surveillance system. *IEEE Transactions on Circuits and Systems for Video Technology*. To appear.

Gabel, M., Renshaw, E., Schuster, A., and Gilad-Bachrach, R. (2012). Full body gait analysis with Kinect. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Gálai, B. and Benedek, C. (2015). Feature selection for lidar-based gait recognition. In *International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, pages 1–5.

Gross, R. and Shi, J. (2001). The CMU Motion of Body (MoBo) Database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA.

Han, J. and Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(2):316–322.

Hofmann, M., Bachmann, S., and Rigoll, G. (2012). 2.5D gait biometrics using the depth gradient histogram energy image. In *Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pages 399–403.

Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., and Rigoll, G. (2014). The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits. *J. Vis. Comun. Image Represent.*, 25(1):195–206.

Kale, A., Cuntoor, N., Yegnanarayana, B., Rajagopalan, A., and Chellappa, R. (2003). Gait analysis for human identification. In *Audio- and Video-Based Biometric Person Authentication*, volume 2688 of *Lecture Notes in Computer Science*, pages 706–714. Springer.

Makihara, Y., Mannami, H., Tsuji, A., Hossain, M., Sugiura, K., Mori, A., and Yagi, Y. (2012). The OU-ISIR gait database comprising the treadmill dataset. *IPSJ Trans. on Computer Vision and Applications*, 4:53–62.

Murray, M. P. (1967). Gait as a total pattern of movement. *American Journal of Physical Medicine*, 46(1):290–333.

Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., and Yagi, Y. (2016). Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 International Conference on Biometrics (ICB)*, pages 1–8.

Tang, J., Luo, J., Tjahjadi, T., and Gao, Y. (2014). 2.5D multi-view gait recognition based on point cloud registration. *Sensors*, 14(4):6124–6143.

Whytock, T., Belyaev, A., and Robertson, N. (2014). Dynamic distance-based shape features for gait recognition. *Journal of Mathematical Imaging and Vision*, pages 1–13.

Wolf, T., Babaee, M., and Rigoll, G. (2016). Multi-view gait recognition using 3d convolutional neural networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 4165–4169.

Yam, C.-Y. and Nixon, M. S. (2009). *Gait Recognition, Model-Based*, pages 633–639. Springer US, Boston, MA.

Zhang, Z., Hu, M., and Wang, Y. (2011). A survey of advances in biometric gait recognition. In *Biometric Recognition*, volume 7098 of *Lecture Notes in Computer Science*, pages 150–158. Springer Berlin Heidelberg.

Zheng, S., Zhang, J., Huang, K., He, R., and Tan, T. (2011). Robust view transformation model for gait recognition. In *International Conference on Image Processing (ICIP)*.