Lidar-based Gait Analysis and Activity Recognition in a 4D Surveillance System

Csaba Benedek, Bence Gálai, Balázs Nagy and Zsolt Jankó

***Abstract*—This paper presents new approaches for gait and activity analysis based on data streams of a Rotating Multi Beam (RMB) Lidar sensor. The proposed algorithms are embedded into an integrated 4D vision and visualization system, which is able to *analyze* and interactively *display* real scenarios in natural outdoor environments with walking pedestrians. The main focus of the investigations are gait based person re-identification during tracking, and recognition of specific activity patterns such as bending, waving, making phone calls and checking the time looking at wristwatches. The descriptors for training and recognition are observed and extracted from realistic outdoor surveillance scenarios, where multiple pedestrians are walking in the field of interest following possibly intersecting trajectories, thus the observations might often be affected by occlusions or background noise. Since there is no public database available for such scenarios, we created and published a new Lidar-based outdoors gait and activity dataset on our website, that contains point cloud sequences of 28 different persons extracted and aggregated from 35 minutes-long measurements. The presented results confirm that both efficient gait-based identification and activity recognition is achievable in the sparse point clouds of a single RMB Lidar sensor. After extracting the people trajectories, we synthesized a free-viewpoint video, where moving avatar models follow the trajectories of the observed pedestrians in real time, ensuring that the leg movements of the animated avatars are synchronized with the real gait cycles observed in the Lidar stream.**

***Index Terms*—multi-beam Lidar, gait recognition, activity recognition, 4D reconstruction**

I. Introduction

The analysis of dynamic 3D (i.e. 4D) scenarios with multiple moving pedestrians has received great interest in various application fields, such as intelligent surveillance [1], video communication or augmented reality. A complex visual scene interpretation system implements several steps starting with people detection, followed by localization and tracking, trying to achieve higher level activity recognition or abnormal event detection functions, and efficient visualization.

Manuscript received March 3, 2016; revised May 22, 2016; accepted June 13, 2016; date of current version June 14, 2016. The work of C. Benedek was supported in part by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. This paper was recommended by Associate Editor Dr. Dong Xu.

Csaba Benedek is with the Distributed Events Analysis Research Laboratory, Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA–SZTAKI), H-1111 Kende u. 13-17 Budapest, Hungary and with the Péter Pázmány Catholic University, H-1083, Práter utca 50/A, Budapest, Hungary. E-mail:benedek.csaba@sztaki.mta.hu

Bence Gálai, Balázs Nagy and Zsolt Jankó are with the Distributed Events Analysis Research Laboratory, MTA–SZTAKI, H-1111 Kende u. 13-17 Budapest, Hungary. E-mail:lastname.firstname@sztaki.mta.hu

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2595331

A critical issue in surveillance of people is the assignment of broken trajectory segments during the tracking process, that are usually produced by frequent occlusions between the people in the scene, or simply by the fact that the pedestrians may temporarily leave the Field of View (FoV). People re-identification [2] requires the extraction of biometric descriptors, which in our case may be *weak features*, since we are focusing on a relatively small number of people (i.e., we are not trying to identify specific people from large databases). On the other hand, in our scenarios the people are non-cooperative, they have to be recognized during their natural behavior, and the process should be (nearly) real time.

Gait as a biometric feature has been extensively examined in the recent decades [3], [4], [5], since psychological experiments already proved in the 1960s that many people can efficiently recognize their acquaintances based on the way they walk [6]. A video-based gait recognition module may be integrated into surveillance systems in a straightforward way, since it does not need additional instrumentation, and it does not require the people to have contact with any special equipment: they may naturally walk in the FoV of the cameras. Although several studies on gait based person identification have been published in the literature (see Sec. I-A for an overview), most existing techniques have been validated in strongly controlled environments, where the gait prints of the test subjects have been independently recorded one after another, and the assignment has been conducted as an offline process. On the other hand, in a realistic surveillance scenario, the gait features should be observed in an arbitrary *wild* (urban or natural) scene, where multiple pedestrians are concurrently present in the field, and they may partially occlude each other. To preserve the online analyzing capabilities of the system, the person assignment should also be performed during the action, where the relative frequency of newly appearing and re-appearing people is arbitrary and unknown.

Apart from person identification, further challenges in 4D scenes are related to the need for efficient visualization of the measurements. By simultaneously displaying different camera views, the observers may monitor large environments, where objects occluded in certain views may be analyzed from other viewpoints. However the efficiency of such a configuration quickly deteriorates with the increase of displays providing unstructured scene information, which is hard to follow for human operators Obtaining realistic 4D video flows of real world scenarios may result in a significantly improved visual experience for the observer compared to watching conventional video streams, since a reconstructed 4D scene can be viewed and analyzed from an arbitrary viewpoint, and virtually

modified by the user. Although there exist stereo vision based solutions for capturing and reconstructing dynamic outdoor scenes, such us [7], they are not fully automatic and they are extremely computation-intensive.

The *integrated*4D (i4D) vision and visualization system proposed in [8] offers a reconstruction framework for dynamic 3D scenarios by integrating two different types of data: outdoor 4D point cloud sequences recorded by a rotating multi-beam (RMB) Lidar sensor, and 4D models of moving actors obtained in an indoor 4D Reconstruction Studio [9] in advance (offline). The system is able to automatically detect and track multiple moving pedestrians in the field of interest, and it provides as output a geometrically reconstructed and textured scene with moving 4D studio avatars, which follow in real time the trajectories of the observed pedestrians. As shown in the workflow of Fig. 1, the RMB Lidar sensor monitors the scene from a fixed position and provides a dynamic point cloud sequence. The measurement is processed to build a 3D model of the static part of the scene and detect and track the moving people. Each pedestrian is represented by a sparse, moving point cluster and a trajectory. A sparse cluster is then replaced with an avatar created in a 4D studio [9]. Finally the integrated 4D scene model can be displayed from an arbitrary user viewpoint.

The basic i4D system [8] had a few notable limitations. *First* only a short time tracking process was implemented, therefore after loosing the trajectories the re-appearing people were always marked as new persons. This re-identification issue has been partially addressed in [10], based on Lidar based weak biometric identifiers featuring the measured height and the intensity histogram of the people's point cloud segments. However, the previous two descriptors may confuse the targets, if their heights and clothes are similar. The vertical resolution of an RMB Lidar sensor is quite low (0.4° in case of the Velodyne HDL 64-E sensor applied in [10]), which means that one needs a height difference of at least 6-8cm between two people for reliable discrimination. Another problem is that the intensity channel of the considered sensor is not calibrated, i.e. the measured intensity values are not necessarily characteristic for a given clothing material, and they may depend on the sensor's distance and the view angle.

The *second* limitation of the [8] system was that although the avatars followed the real person trajectories, always turning according to the trajectories' tangents, the animated leg movements were not synchronized with the real walk cycles. The step cycles recorded in the 4D Studio were simply repeated continuously disregarding the step frequency and phase information, having a distracting visual impact. Therefore, gait analysis may also contribute to the improvement of the existing animation framework, by continuously extracting actual gait phases from the Lidar measurements and using the extracted phase information for realistic animation of the walking models. The contributions of the present paper focus in part on overcoming the above mentioned limitations, by supporting the re-identification and animation steps of the system with gait-based features.

A *third* limitation of [8] that we try to overcome in this paper is that previously only "normal" walking scenarios were considered. However, in a real surveillance environment one should expect to observe various activities such as people making phone calls, bending, checking their watches, waving their hands, etc. Since in the 4D studio the previously mentioned motion templates can be recorded in a straightforward way for the free-viewpoint video output, the main challenge here is to implement an automatic activity recognition module based on the Lidar point cloud sequence.

A. Related work in gait analysis

In this section, we give an overview on existing visual gait analysis and recognition techniques from the literature, focusing on their connections to our measurement scenarios.

Several methods tackle the detection problem on videos of monocular optical cameras. Since we can only assume the subjects' side view visibility in very specific environments [11], [12], a key research issue is to find a view invariant representation of the extracted gait features. Among various approaches, a view transformation model using a multi-layer perceptron is introduced in [13], while the gait energy image (GEI) representation has been adopted in [14], [15]. A new dimensionality reduction technique is presented for the average silhouettes in [16]. Patch Distribution Features are built on the GEI representation in [17], [18]. A new image-to-class distance metrics was proposed in [19] to enable efficient comparison of different gait patterns. [20] performs spatio-temporal silhouette print comparison via the Dynamic Time Warping (DTW) signal processing algorithm, and features from simple silhouette averaging are utilized in [21]. A number of techniques transform the objects into a canonical shape representation [22], [23]. All the above methods aim to maximize the detection performance for different public multi-view gait databases [24], such as the *CASIA gait dataset* [25], the *USF database* [26], and *the CMU Motion of Body (MoBo)* Database [27]. Although these datasets contain motion sequences of many pedestrians from different viewpoints for cross view validation, they are recorded in strongly controlled indoor or outdoor environments in terms of illumination, background surfaces and background motions. Additionally, the test subjects follow fixed trajectories [26] or walk on a treadmill [27], conditions which impose significant restrictions versus real surveillance scenarios, where the targets may move arbitrarily. The HID-UMD database [28] contains walk videos captured in more general outdoor environments, with various view angles, camera distances and background parameters. However, similarly to the other mentioned datasets, the pedestrians are walking alone in each video sequence, a constraint which makes high quality silhouette extraction a feasible task. On the other hand, in real dynamic scenes with a large FoV, we must expect multiple freely walking pedestrians, possibly occluding each other, therefore the critical silhouette extraction step might become a bottleneck for the whole process. Problems caused by occlusions can be partially handled by information fusion of different views [29], however this approach requires a carefully positioned and calibrated multi-camera system, making quick temporary installation difficult for applications monitoring ad-hoc or special events.

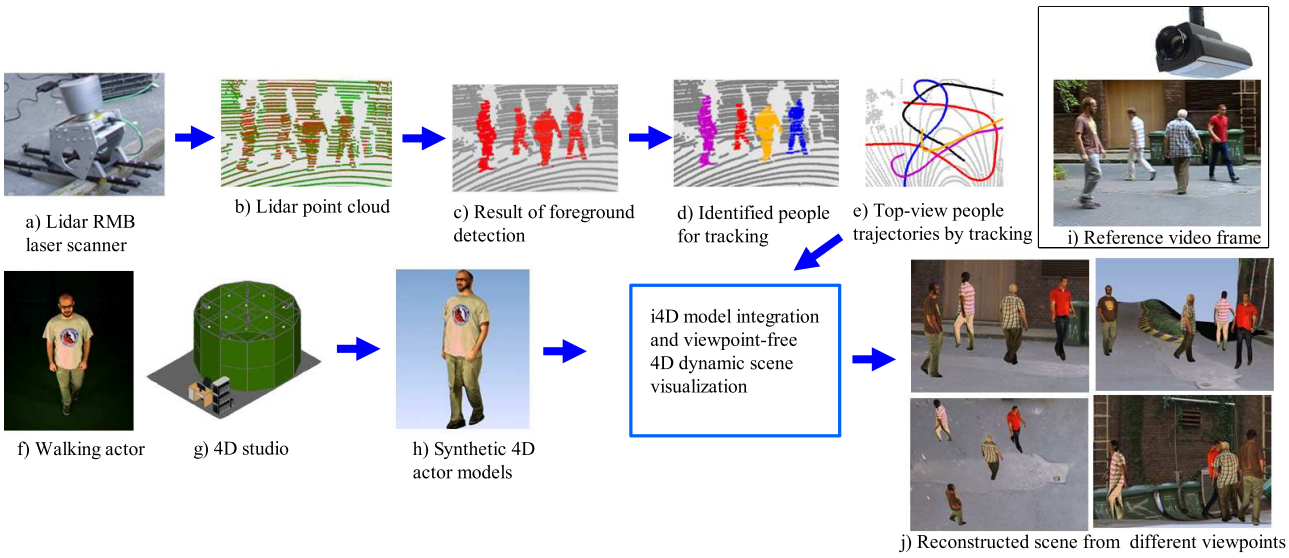


Fig. 1. Workflow of the i4D system framework [8]. Figures a)-e) demonstrate the steps of Lidar based moving object detection and multi-target tracking, f)-h) shows the reconstruction process of moving avatars in the 4D studio, i) displays a reference video image from the same scenario (not used by the i4D workflow) and j) is a snapshot from the reconstructed 4D scenario, shown from four different viewpoints

A possible option for obtaining depth information from the scene is using stereo cameras or Time-of-Flight (ToF) technologies. Cheap Kinect sensors have been investigated for gait analysis in a number of works [30], [31], [32], and a corresponding gait database has already been published [33] for reference. However Kinects are still less efficient for applications for real life outdoor scenarios due to their small FoV and range (resolvable depth is between 0.8m – 4.0m), and the low quality outdoor performance of the sensor, especially in direct sunlight. A 2D laser range scanner has been used to measure gait characteristic in [34], [35], however, due to its 2D nature, the scanner is not able to perform object classification, and it cannot oversee large scenes with multiple objects either.

Velodyne's Rotating Multi-Beam (RMB) Lidar sensor is able to provide point cloud sequences from large outdoor scenes with a frame-rate of 15 Hz, with a 360° FoV, and produces point clouds with approximately 65K points/frame with a maximum radius of 120m. The RMB Lidar sensor does not need any installation or calibration after being placed into a new environment. However, the spatial density of the point cloud is quite sparse, showing a significant drop in the sampling density at larger distances from the sensor, and we can also see a ring pattern with points in the same ring much closer to each other than points of different rings. According to our measurements, the the size of a point cloud associated to a person in a courtyard with a radius of 10-20m varies between 0.18–0.5K points, which is *two orders of magnitude* smaller than the figures of Kinect (10-20K points/person), and also significantly lower than the density of the stereo camera measurements from [36].

B. The contributions of the paper

In this paper, we investigate the possibility of using a RMB Lidar sensor (specifically, the Velodyne HDL 64-E) for visual gait analysis and activity recognition, supporting pedestrian re-identification and 4D visualization tasks in realistic surveillance environments. Although pedestrian detection and tracking tasks have already been conducted on RMB Lidar measurements [10], [37], to our best knowledge our research [38], [39] has been the first attempt to involve such sensors in gait recognition. Due to the low spatial resolution of the sensor, and the presence of partially incomplete pedestrian shapes due to various occlusion effects, we decided to follow a *model free* approach, in contrast to *model based* methods [40], [41], [42] which fit structural body part models to the detected objects and extract various joint angles or body segment length parameters. For example, [41] used particle swarm optimization for model fitting based on the edge distance map, which definitely requires high quality silhouettes. On the contrary, our main efforts focus on noise tolerant robust extraction of the descriptors and the integration of the efficient fusion of the gait parameters with other feature modalities.

II. LIDAR BASED SURVEILLANCE FRAMEWORK

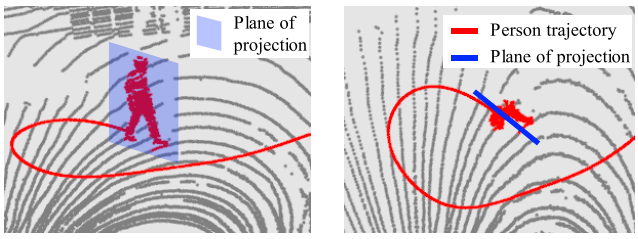
The main steps of the processing pipeline are demonstrated in Fig. 1. The RMB Lidar records 360° range data sequences of irregular point clouds (Fig. 1(b)). To separate dynamic foreground from static background in a range data sequence, a probabilistic approach [43] is applied. To ensure real-time operation, we project the irregular point cloud to a cylinder surface yielding a depth image on a regular lattice, and perform the segmentation in the 2D range image domain. We model the statistics of the range values observed at each pixel position as a Mixture of Gaussians and update the parameters similarly to the standard approach [44]. The background is modeled by the Gaussian components with the highest weight values in the mixture, and outlier detection enables the extraction of the possible motion regions. However, by adopting the above scheme, we must expect several spurious effects, caused by

the quantisation error of the discretised view angle and background flickering, e.g., due to vegetation motion. These effects are significantly decreased by a dynamic MRF model [43], which describes the background and foreground classes by both spatial and temporal features. Since the MRF model is defined in the range image space, the 2D image segmentation must be followed by a 3D point classification step by resolving the ambiguities of the 3D-2D mapping with local spatial filtering. Using a contextual foreground model, we remove a large part of the irrelevant background motion which is mainly caused by moving tree crowns. A sample frame for the result of foreground detection is shown in Fig. 1(c).

The next step is pedestrian detection and tracking. The input of this component is the RMB Lidar point cloud sequence, where each point is marked with a segmentation label of foreground or background, while the output consists of clusters of foreground regions so that the points corresponding to the same object receive the same label over the sequence (Fig. 1(d)).

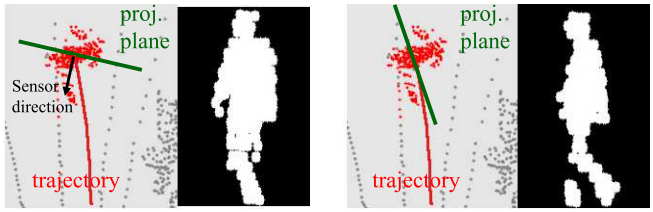
First, the point cloud regions classified as foreground are clustered to obtain separate blobs for each moving person candidate. A regular lattice is fit to the ground plane and the foreground regions are projected onto this lattice. Morphological filters are applied in the image plane to obtain spatially connected blobs for different persons. Then the system extracts appropriately sized connected components that satisfy area constraints determined by lower and higher thresholds. The centre of each extracted blob is considered as a candidate for foot position in the ground plane. Note that connected pedestrian shapes may be merged into one blob, while blobs of partially occluded persons may be missed or broken into several parts. Instead of proposing various heuristic rules to eliminate these artifacts at the level of the individual time frames, a robust multi-tracking module has been developed, which efficiently handles the problems at the sequence level.

The pedestrian tracking module combines Short-Term Assignment (STA) and Long-Term Assignment (LTA) steps. The STA part attempts to match each actually detected object candidate with the current object trajectories maintained by the tracker, by purely considering the projected 2D centroid positions of the target. The STA process should also be able to continue a given trajectory if the detector misses the concerning object for a few frames due to occlusion. In these cases the temporal discontinuities of the tracks must be filled with estimated position values. On the other hand, the LTA module is responsible for extracting discriminative features for the re-identification of objects lost by STA due to occlusion in many consecutive frames or leaving the FoV. For this reason, lost objects are registered to an archived object list, which is periodically checked by the LTA process. LTA must also recognize when a new, previously not registered person appears in the scene. Finally, we generate a 2D trajectory of each pedestrian. Even with applying Kalman filtering, the extracted 2D raw object tracks proved to be quite noisy, therefore, we applied a 80% compression of the curves in the Fourier descriptor space [45], which yielded the smoothed tracks displayed in Fig. 1(e).



(a) Bird's view (b) Top view

Fig. 2. Silhouette projection: (a) a tracked person and its projection plane in the point cloud from bird's view (b) projection plane from top view, taken as the tangent of the smoothed person trajectory.



(a) Circular tangent (b) Trajectory tangent (sideview)

Fig. 3. Silhouette projection types from top-view. (a) the projection plane's normal points towards the sensor (undefined silhouette orientation) (b) the projection plane is the tangent of the trajectory (sideview silhouettes)

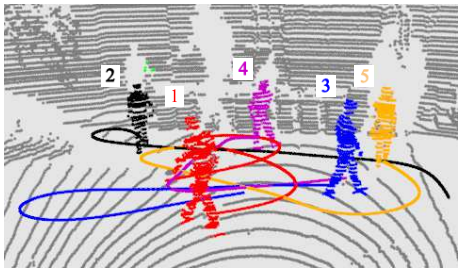
III. Lidar based Gait analysis

In the proposed framework, the main goal of gait investigation is to support the long-term assignment (LTA) process of the tracking module. To fulfill the requirements of real surveillance systems, we need to extract unique biometric features online during the multi-target tracking process from the measurement sequence.

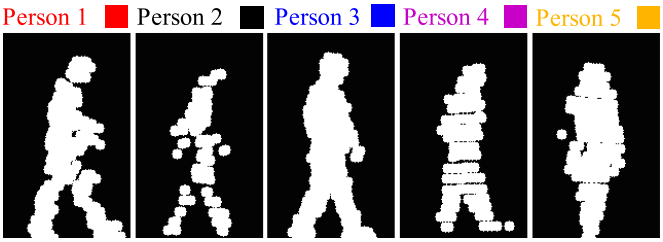
For gait analysis, we focus on 2D silhouette based approaches, which are considered quite robust against low resolution and partial occlusion artifacts, due to capturing information from the whole body. The first step is projecting the 3D points of a person in the RMB Lidar point cloud to an appropriately selected image plane. Since the FoV of the Velodyne sensor is circular, a straightforward projection plane could be taken at a given ground position as the local tangent of the circle around the sensor location (see Fig. 3(a)). However this choice would not ensure viewpoint invariant features as the silhouette's orientation may be arbitrary.

(a) MoBo (image) (b) Lidar - near (c) Lidar - far

Fig. 4. Comparison of the (a) high resolution CMU MoBo silhouettes captured with an optical video camera and (b)-(c) our low quality RMB Lidar based silhouettes



(a) Pedestrians on the Lidar frame

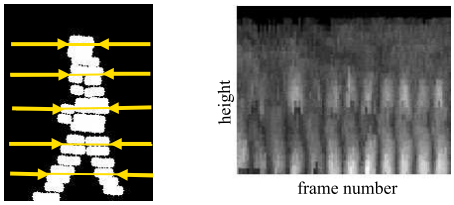


(b) Projected silhouettes

Fig. 5. Comparison of the (a) Output of the multi-pedestrian tracker for a sample Lidar frame (person point clouds+trajectories)(b) projected pedestrian silhouettes on the selected Lidar frame

Instead, we interpolate the side view projections of the 3D human silhouettes, by exploiting the assumption that people mostly walk forwards in the scene, turning towards the tangent direction of the trajectory. At each time frame, we project the point cloud segment of each person to the plane, which intersects the actual ground position, is perpendicular to the local ground plane, and it is parallel to the local tangent vector of the Fourier-smoothed trajectory from the top view (Fig. 2 and 3(b)).

The projected point cloud consists of a number of separated points in the image plane, which can be transformed into connected 2D foreground regions by morphological operations. In Fig. 4 (a) and (b), we can compare a (spatially) downscaled silhouette from the CMU (MoBo) Database [27], and a *quite clean* silhouette provided by our system. We can see that due to the morphological dilation kernels the Lidar-based masks retain much less detail of the the object contour, but the shape is clearly observable at a coarse scale. In addition, a main advantage is of the Lidar technology is that the laser measurement is directly obtained in the 3D Euclidean coordinate space, without perspective distortion and scaling effects, thus the projected silhouettes may be also compared without re-scaling. However, the density of the point cloud representing a given person is significantly lower at a larger distance from the sensor, yielding silhouettes which have discontinuities as demonstrated in Fig. 4(c). Further challenging samples can be observed in Fig. 5, which shows a snapshot from a 5-person-sequence with the actually extracted silhouette masks. *First*, the silhouettes of Persons 2 and 4 are disconnected, since they are far away from the sensor. *Second*, for people walking towards the sensor, the 2.5D measurement provides a frontal or back view, where the legs may be partially occluded by each other (see Person 5). *Third*, some silhouette parts may be occluded by other people or field objects in a realistic surveillance scene (see Person 2).



(a) Width vector (b) Silhouette print

Fig. 6. Features of the silhouette print [20] technique on Lidar data

In the following parts of this section, we analyze various features for gait-based person re-identification in the projected person-image sequences. For comparison, we implemented four different model-free silhouette or range image based approaches in our Lidar-based surveillance framework. The first three techniques are Lidar-focused modifications of state-of-the-art approaches, proposed earlier for standard optical and Kinect data, while the fourth method is an improvement of our model from [38]. By each selected method, we had to explore first whether their expected input feature maps can be derived from the RMB Lidar streams. Usually the adoptions did not proved to be straightforward, and we experienced that the above mentioned limitations of silhouette extraction (especially occlusion and low resolution) significantly affected the performance.

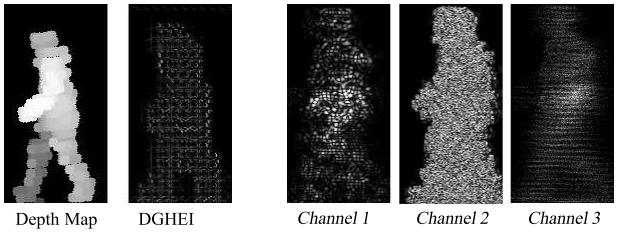
A. Silhouette print

Kale et al. [20] used the width of the outer contour of a binarized silhouette as the basic feature. In this method, a bounding box is placed around the extracted silhouette patch, which is divided into D equal box-parts along the vertical axis. Then the width of the silhouette is stored in each box-part, yielding a D dimensional (used $D = 20$) width-vector at a given time frame (Fig. 6(a)). The width-vectors of consecutive frames are combined into an image called silhouette print (SP) image, which is visualized in Fig. 6(b), in which brighter pixels refer to larger values of the width vectors. Similarities between the prints are calculated using the dynamic time warping (DTW) algorithm [20].

Before starting the evaluation in our Lidar dataset, we validated our implementation on the original CMU MoBo [27] (optical) database, and reproduced similar efficient results to [20]. Thereafter, the adaptation of the method to the more challenging Lidar-scene has been straightforward: we generated 5 prints for every person for gallery (training) data, and during the re-identification step we have chosen the person, whose galleries showed in average the lowest DTW distance from the current probe (test) data.

B. Depth Gradient Histogram Energy Image

The *Depth Gradient Histogram Energy Image* (DGHEI) technique was proposed for gait recognition in Kinect sensor data [32]. Instead of binarized silhouettes, the inputs of DGHEI are depth images derived from the 2.5D measurements. Depth gradients are calculated with histogram binning,



(a) DGHEI features (b) CGCI feature maps

Fig. 7. Feature maps of (a) DGHEI [32] and (b) CGCI [30] on Lidar data

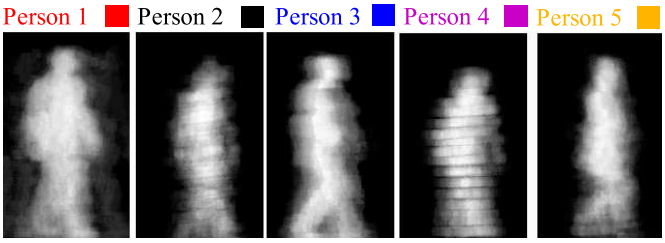


Fig. 8. Lidar based Gait Energy Images extracted for the people of Fig. 5

and the histogram bins are averaged for full gait cycles. Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) are used for dimension reduction, and a nearest neighbor classifier for classification. We have implemented this workflow for the RMB sensor measurements: a sample depth image and the corresponding DGHEI feature map is shown in Fig. 7(a).

C. Color Gait Curvature Image

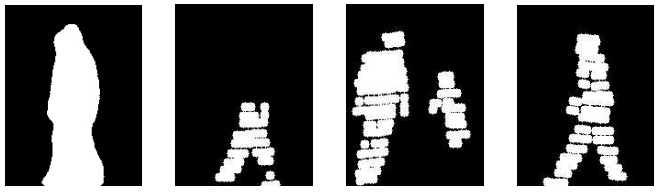
The *Color Gait Curvature Image* (CGCI) approach has also been introduced for Kinect point clouds [30]. This technique uses three 2.5D gait features: Gaussian curvature, mean curvature and local point density, which are combined into a 3-channel descriptor map, shown in Fig. 7(b). Then, 2D Discrete Cosine Transform and 2D-PCA steps are applied to the feature channels separately. Classification is performed by calculating a weighted sum of the absolute differences of the three feature components.

D. Proposed Gait Energy Image based Approach

In our proposed model, we adopt the idea of Gait Energy Image (GEI) based person recognition to the Lidar surveillance environment. The original GEI approach was introduced by Han and Bhanu in 2006 [14] for conventional optical video sequences. GEIs are derived by averaging the binary person silhouettes over the gait cycles:

$$G(x,y) = \frac{1}{T}\sum_{t=1}^{T} B_t(x,y) \qquad (1)$$

where $B_t(x,y) \in \{0,1\}$ is the (binary) silhouette value of pixel (x,y) on time frame t, and $G(x,y) \in [0,1]$ is the (rational) GEI value. In [14] a person was represented by a set of different GEI images corresponding to the different observed gait cycles, which were compressed by PCA and MDA.



(a) Filter mask (b) Dropped1 (c) Dropped2 (d) Kept frame

Fig. 9. Demonstration of the automatic frame selection step.

Thereafter person recognition was achieved by comparing the gallery (training) and probe (test) features.

In our environment, a number of key differences had to be implemented compared to the reference model [14], leading to a new descriptor that we call *Lidar-based Gait Energy Image* (LGEI). The first key contribution is, that since the RMB Lidar measurement sequences have a significantly lower temporal resolution (15 fps), than the standard video flows ($\geq$25 fps), samples from a single gait cycle provide too sparse information. For this reason, we do not separate the individual gait cycles before gait print generation, but we select k (used $k = 100$) random *seed frames* from each person's recorded observation sequence instead, and for each *seed* we average the l consecutive frames (used $l = 60$) to obtain a given LGEI sample. This way, k LGEIs are generated for each individual, and to enable later data compression, global PCA and MDA transforms are calculated for the whole dataset.

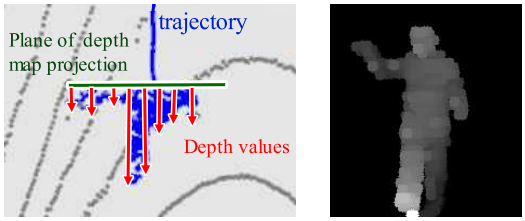
The second key difference is, that instead of following the direct GEI-set based person representation and vector comparison of [14], we propose here a neural network based approach. Similarly to [46] we have chosen to use a committee of a Multi-Layer Perceptron (MLP) and a convolutional neural network (CNN), both having N outputs, where N is equal to the number of people in the training scenario. The dominant 35 PCA and 5 MDA components of the LGEIs are used to train a Multi Layer Perceptron (MLP) for each person, while the CNN inputs are the raw 2D LGEIs. We used *tanh* activation functions whose output is in the $[-1,1]$ domain. Thus for a training sample of the ith person, the ith network's prescribed output value is 1, while the remaining outputs are -1.

In the person recognition phase, we generate probe LGEIs for each detected and tracked subject: we start from a random seed frame of the sequence and average the upcoming l consecutive silhouettes. The trained networks produce outputs within the range $o_{\text{MLP}}, o_{\text{CNN}} \in [-1,1]$, and the ith output (corresponding to the ith trained person) of the MLP-CNN committee is taken as the maximum of the outputs of the two networks: $o^i = \max(o^i_{\text{MLP}}, o^i_{\text{CNN}})$, $i = 1,\ldots,N$. As a valid identification of a given G probe LGEI, only positive $o^i(G)$ values are accepted. Therefore, with the notation of $i_{\max} = \text{argmax}_i\, o^i(G)$, sample G recognized as person $i_{\max}$, if $o^{i_{\max}} > 0$, otherwise we mark G as *unrecognized*.

For reducing further artifacts caused by frequent occlusions, we also developed a frame selection algorithm for our LGEI-based approach. A binary mask is created by summing and thresholding the consecutive silhouettes for every person (Fig. 9(a)). For every silhouette we calculate its internal and external

(a) Video frame (only reference) (b) Lidar frame

Fig. 10. A sample frame from an outdoor test sequence used for activity recognition



(a) Projection (b) Depth map

Fig. 11. Demonstration of the the frontal projection and depth map calculation for activity recognition. Projection plane is perpendicular to the trajectory.

area w.r.t. the mask. If the internal area is less then 40% of the mask's area (Fig. 9(b)) or the external area is more then 30% of the mask's area (Fig. 9(c)) the frame is discarded from the LGEI calculation. As a result, several irrelevant frames get dropped and do not compromise the LGEI calculation. An example for a kept silhouette frame is shown in Fig. 9(d). Note that the above sample collection scheme is is not effected by prior gait cycle estimation in contrast with [14].

IV. Action Recognition

While the previous section analyzed gait as biometric feature during *normal walk*, the analysis can lead to wrong results if the input sequence does not only contain forward-walking people, but other actions as well. On the other hand, the recognition of various actions can provide valuable information in surveillance systems. The main goal of this section is to propose features for recognizing selected - usually rarely occurring - activities in the Lidar surveillance framework, which can be used for generating automatic warnings in case of specific events, and removing various 'non-walk' segments from the training/test data of the gait recognition module.

In the literature, one can find a number of activity recognition approaches based on image sequences, point clouds or depth maps, where occupancy patterns are calculated [47] or different features are extracted such as spatio-temporal context distribution of interest points [48], histogram of oriented principal components [49] or oriented 4D normals [50], and 3D flow estimation [51]. However, the sparsity of Lidar point clouds (versus Kinect) becomes a bottleneck for extracting the above features. Based on experiments with various descriptors, we decided to follow a map-averaging approach again, which is detailed in the following section. Apart from normal walk, we have selected five events for recognition: *bend*, check *watch*, *phone* call, *wave* and wave two-handed (*wave2*) actions. A sample outdoor frame with four people is shown in Fig. 10.

A. Selected features

Our approach for action recognition is motivated by the LGEI based gait analysis technique (Sec. III-D), however, various key differences have been implemented here.

First, while gait could be efficiently analyzed from side-view point cloud projections, the actions listed above are better observable from a frontal point of view. For this reason, we have chosen a projection plane for action recognition, which is perpendicular to the local trajectory tangent, as demonstrated in Fig. 11(a). (Note that this plane is also perpendicular to the LGEI's projection plane).

Second, various actions, such as waving or making phone calls produce characteristic local depth texture-patterns (e.g. the hand goes forward for waving). Therefore, instead of deriving binarized silhouettes (Fig. 4), we create depth maps by calculating the point distances from the projection plane according to Fig. 11(a), a step which yields a depth image shown in Fig. 11(b). Then, we introduce the *averaged depth map* (ADM) feature as a straightforward adoption of the LGEI concept, so that we average the depth maps for the last τ frames, where τ is the a preliminary fixed time window related to the expected duration of the activities (we used $\tau = 40$ frames uniformly). ADM sample images for each activity are shown in Fig. 12 (top row).

Third, while gait is considered a low-frequency periodic motion of the whole body, where we do not lose a significant amount of information by averaging the consecutive images, the above actions are aperiodic and only locally specific for given body parts. For example, waving contains sudden movements, which yield large differences in the upper body regions of the consecutive frames. Thus, apart from ADM we introduce a second feature, called *averaged XOR image* (AXOR), which aims to encode information about the motion dynamics. An exclusive-OR (XOR) operation is applied on two consecutive binarized frontal silhouettes, and the AXOR map is calculated by averaging these binary XOR images and taking the squares of the average values. The AXOR map displays high values for the regions of sudden movements, as shown in Fig. 12 (bottom row), especially regarding the waving actions in images (e) and (f).

B. Training and recognition

For each action from the set *bend*, *watch*, *phone*, *wave* and *wave2*, two separate convolutional neural networks (CNN) were trained, one for the ADM and one for the AXOR features, respectively. As explained in [52], a small (4-layer) CNN could be constructed, using the spatially downscaled (to 20×16 pixels) and normalized ADM and AXOR feature maps. During the training of the CNNs, we prescribed the output values 1.0 for positive and -1.0 for negative samples by each activity. The negative training data also included various samples from normal *walking*. The outputs of the CNNs range from -1.0 to 1.0 , and a probe sample is recognized as a given action if the corresponding ADM-based and AXOR-based CNN outputs

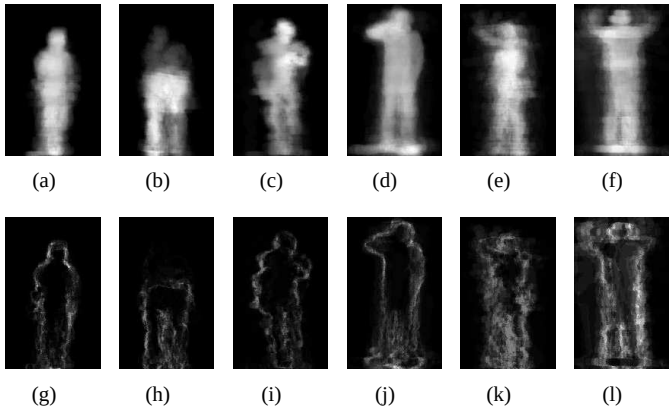


Fig. 12. ADM (top row) and AXOR (bottom row) for the (a,g) walk, (b,h) bend, (c,i) check watch, (d,j) phone call, (e,k) wave and (f,l) wave two-handed (wave2) actions.

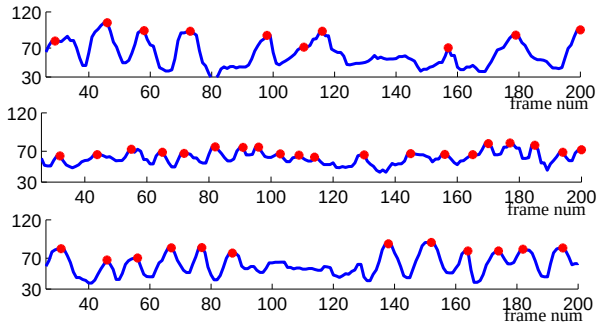


Fig. 13. Silhouette width sequences for three selected persons from a test scenario - used for gait step synchronization during visualization

both surpass a ν decision threshold (used $\nu = 0.6$). If no activity is detected, we assume that the observed person is in the *walking* state. If multiple CNN outputs surpass the decision threshold, we select the action with the highest confidence.

V. 4D SCENE VISUALIZATION

The visualization module takes as input the trajectories of the identified walking people, and the timestamp and location of the recognized actions. As output a free-viewpoint is synthesized, where moving animated avatars follow the motions of the observed people. The moving avatars are properly detailed, textured dynamic models which are created in a 4D reconstruction studio, whose hardware and software components are described in [9]. The 4D person models can be placed into an arbitrary 3D background (point cloud, mesh, or textured mesh), which can be either created manually with a CAD system, or by automatic environment mapping of the Lidar measurements [8].

The last step of the workflow is the integration of the system components and visualization of the integrated model. The walking pedestrian models are placed into the reconstructed environment so that the center point of the feet follows the trajectory extracted from the Lidar point cloud sequence. The temporal synchronization of the observed and animated leg movements is implemented using the gait analysis. This step requires an approximation of the gait cycles from the Lidar measurement sequence, however the accuracy is not critical here, but the viewer has to be visually convinced that the leg movements are correct. The cycle estimation is implemented by examining the time sequence of the 2D bounding boxes, so that the box is only fitted to the lower third segment of the silhouette. After a median filter based noise reduction, the local maxima of the bounding box width sequence are extracted, and the gait phases between the key frames are interpolated during the animation. Although as shown in Fig. 13, the width sequences are often notably noisy, we experienced that the synthesized videos provide realistic walk dynamics for the users.

VI. DATASET FOR EVALUATION

Utilizing relevant test data is a key point in evaluation. Since to our best knowledge no Lidar based gait or activity recognition dataset has been published yet for surveillance environments, we have created the SZTAKI Lidar Gait-and-Activity (SZTAKI-LGA) database[1], which is designed for the evaluation of gait based person identification and activity recognition in a multi-pedestrian environment.

For gait analysis, our proposed SZTAKI-LGA database contains *ten* outdoor sequences captured in a courtyard by a Velodyne HDL 64-E RMB Lidar sensor. All the sequences have 15 fps frame rate, their length varies between 79 and 210 seconds (in average 150 sec.), and each contains 3-8 people walking simultaneously in the scene. In each case, the test subjects were asked to walk naturally in the scene, then all leave the Field of View, re-appear in a different order, and walk till the end of the sequence. This *screen-play* enables to test gait descriptors in realistic surveillance situations, with the goal of matching the corresponding gait patterns collected in the first (*training*) and second (*probe*) parts of each test scenario. Since the sequences were recorded in different seasons, we can also investigate how different clothing styles (such as winter coats or t-shirts) influence the discriminating performance of the observed gait features.

For *action recognition* purposes we recorded 1 indoor and 9 outdoor sequences with a total time of 633 seconds. The test data contains various examples for the five addressed activities: *bend* (88 samples), *watch* (53), *phone* (50), *wave* (58) and *wave2* (46) which are extracted from the sequence. Each sequence contains multiple pedestrians, and the typical length range of a given annotated activity sample varies between 40-100 frames.

VII. EXPERIMENTS AND DISCUSSION

We have evaluated the proposed gait based biometric identification and activity recognition algorithms on the SZTAKI-LGA database. The structures of the convolutional neural networks used for gait and activity recognition were similar, only the second layer's type, the number of feature maps and the kernel size parameters were different, as detailed in Fig. 15. The MLP component in gait analysis used 6 hidden neurons and N outputs, equal to the number of people in the training scenario.

[1]The SZTAKI-LGA database is available at the following URL: `http://web.eee.sztaki.hu/i4d/SZTAKI-LGA-DB`.

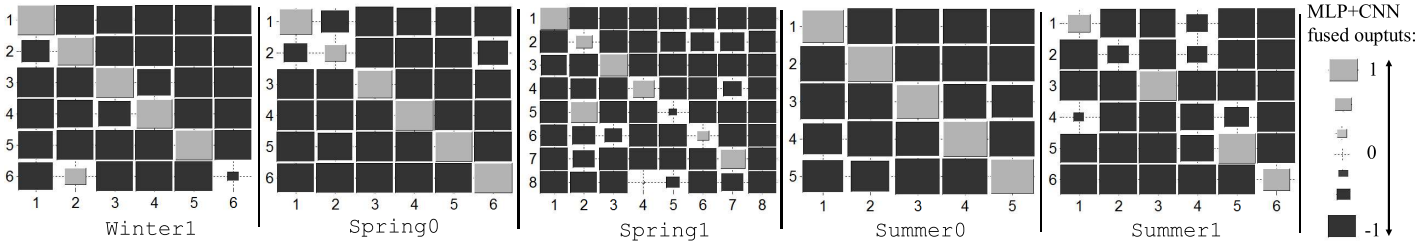
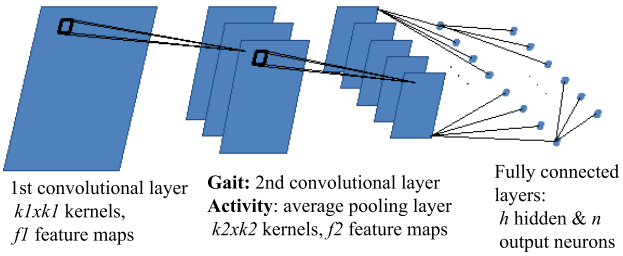


Fig. 14. Quantitative evaluation of LGEI based matching between the gallery (columns) and probe (rows) samples. Rectangles demonstrate the CNN+MLP outputs, the ground truth match is displayed in the main diagonal.



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

Fig. 15. Structure of the used convolutional neural networks (CNN). By gait recognition, N is equal to the number of people in the training set.

TABLE I
EVALUATION RESULTS OF THE COMPARED METHODS: RATES OF CORRECT RE-IDENTIFICATION. N EQUALS THE NUMBER OF PEOPLE.

Scene	N	SP+ DTW	DG- HEI	CGCI	LGEI		
					CNN	MLP	Mix
`Winter0`	4	0.96	0.97	0.36	0.94	0.98	**0.99**
`Winter1`	6	0.33	0.89	0.27	0.85	0.90	**0.95**
`Spring0`	6	0.64	0.81	0.32	0.91	0.95	**0.98**
`Spring1`	8	0.33	0.59	0.20	0.63	0.66	**0.70**
`Summer0`	5	0.39	0.97	0.40	0.99	0.95	**1.00**
`Summer1`	6	0.33	0.83	0.29	0.77	0.95	**0.95**
`Summer2`	3	0.33	0.98	0.53	0.96	0.99	**0.99**
`Summer3`	4	0.50	**0.94**	0.32	0.94	0.93	**0.94**
`Summer4`	4	0.25	**0.95**	0.27	0.91	0.90	0.91
`Summer5`	4	0.50	**0.80**	0.32	0.77	0.74	**0.80**
Average	5	0.46	0.87	0.33	0.87	0.90	**0.92**

TABLE II
PERFORMANCE IMPROVEMENTS CAUSED BY TRAJECTORY BASED PROJECTION PLANE ESTIMATION (TT) AND FRAME SELECTION (SF) USING THE LGEI METHOD.

Scenario	AF+CT	AF+TT	SF+CT	SF+TT
`Winter1`	0.78	0.85	0.81	0.95
`Spring0`	0.80	0.95	0.81	0.98
`Summer0`	0.99	0.99	1.00	1.00
`Summer1`	0.75	0.79	0.83	0.95

Projection plane: circular tangent (CT) or trajectory tangent (TT)
Frames composing LGEIs: all frames (AF) or selected frames (SF)

TABLE III
COMPUTATIONAL TIME (IN SECONDS) OF THE MAIN STEPS OF THE DIFFERENT APPROACHES FOR THE LIDAR BASED GAIT DATABASE

Method	Training set generation	CNN/MLP training	Recognition of 100 test samples
SP+DTW	8.43	-	43.7
DGHEI	110.3	-	0.98
CGCI	108.9	-	0.26
LGEI	142.7	46.9	0.98

A. Evaluation of gait recognition

The gait recognition module has been validated on all the 10 test gait-sequences of the database. We compared the performance of the four different features introduced in Sec. III-A-III-D: *Silhouette print* & Dynamic Time Warping (SP+DTW), *Depth Gradient Histogram Energy Image* (DGHEI), *Color Gait Curvature Image* (CGCI), and *Lidar Based Gait Energy Image* with MLP+CNN committee (LGEI). All the methods (except the silhouette print) were trained using 100 gallery feature maps for each person, extracted from the *training* parts of the sequences. In the evaluation phase, we generated 200 probe maps of each test subject from the *test* segments of the videos. Each probe sample was independently matched to the trained person models, thus we used $200 \cdot N$ test samples in a scenario with N people. For evaluating the performance of the different methods, we calculated the rate of the correct identifications among all test samples, and listed the obtained results in Table I.

Although according to their introducing publications, both the CGCI [30] and DGHEI [32] methods proved to be notably efficient for processing Kinect measurements, their advantages could not be exploited by dealing with the much sparser Velodyne point clouds. In particular, as we can observe in Table I, the CGCI method proved to be the less successful among all the tested techniques for the low density Lidar data, an observation that that could have already been predicted by examining the visually featureless CGCI descriptor channels from Fig. 7(b).

By testing the width-vector based SP+DTW approach [20], we experienced that it only favored the first test scene (`Winter0`), which included nearly complete silhouettes with noiseless contours. However as the quality of silhouettes decreased due to frequent occlusions, and several holes and discontinuities appeared in more crowded tests scenes, the

SP+DTW approach provided quite low recognition rates.

The DGHEI [32] proved to be the second best gait descriptor, outperformed only by our LGEI based method by 5% overall. This observation is not surprising, considering that the DGHEI approach has originally been proposed by extending the Gait Energy image (GEI) with depth gradient extraction and direction histogram aggregation. As detailed in [32] the above improvement increased the performance when high-quality depth images were available, however, in our scenarios with lower resolution depth maps (see Fig. 7(a)) these features could not be efficiently utilized, and the performance become slightly lower than with the LGEI approach. Note that we have also tested the DGHEI based recognition with CNN and MLP neural networks, but this modification did not yield improvement versus the original nearest neighbor classifier proposed in [32].

Our proposed LGEI solution has been tested first by separately using the MLP and CNN networks, and thereafter with the MLP+CNN committee. As the last three columns of Table I confirm, the MLP and CNN outperformed each others on a case-by-case basis, and the committee has generally resulted in improved results over the two network components. As already shown in [39], in LGEI classification the MLP-CNN committee could also outperform the standard Vector Comparison approach proposed in [14].

Table I also demonstrates that compared to the SP+DTW, DGHEI and CGCI techniques the LGEI method provided superior results in most of the test scenarios. The performance drop observed by some of the *more crowded* (6-8 person) scenes has been principally caused by the increased number of occlusions which obviously yielded lower quality input data for the classification framework. As examples, the score matrices between the trained neural networks and the measured gait patterns from the different test subject are displayed in Fig. 14 for five test scenes. This figure highlights the background of the varying performance in the different test cases: from the point of view of (LGEI-based) gait recognition `Spring0` and `Summer0` proved to be *simple* scenarios with nearly *diagonal* score matrices, while `Spring1` and `Summer1` are quite *difficult* sequences, where the measurable benefits of the LGEI technique are the most apparent compared to the weaker performing reference approaches.

By further examination of the LGEI method, we investigated the improvements caused by two auxiliary innovations of our proposed approach:

- Applying trajectory tangent (TT) oriented planes of silhouette projection instead of the straightforward circular tangent (CT) direction (refer to Fig. 3 in Sec. III).
- Automatic selection of frames (SF) instead of using all frames (AF) in LGEI generation, by dropping the presumptively low quality silhouettes (Fig. 9 in Sec. III-D).

As shown in Table II for four selected sequences, both new algorithmic steps yielded notable improvements in the recognition rates.

Our next evaluation stage addresses the performance variation of LGEI based gait recognition, by increasing the number of people in the database. As discussed above, with more than 8 people at the same time in our courtyard we obtained notably degraded silhouette shapes, an artifact caused by the capture conditions, but independent of the biometric separating capability of the gait features. Exploiting that in our 10 test sequences 28 different people have appeared, we collected the silhouette sequences of the different test subjects from all test scenarios into a global database. Then, we selected step by step 2,3,. . .,28 people from the database, and each time we trained and evaluated an LGEI-CNN+MLP committee for the actual subset of the people (using separated training and test samples). The diagram of the observed recognition rates as a function of the number of persons is displayed in Fig. 16, which shows a graceful degradation in performance, staying steadily around 75% for 17-28 people.

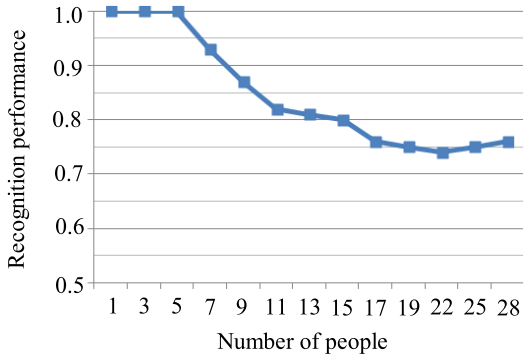


Fig. 16. The overall results on the whole dataset.

TABLE IV
THE CONFUSION MATRIX OF ACTION RECOGNITION.

Detect→ GT↓	*Bend*	*Watch*	*Phone*	*Wave*	*Wave2*	*FN*	*FP*
Bend	85					3	
Watch		37	1		4	11	3
Phone		5	36	2	2	5	6
Wave			4	44	5	5	3
Wave2			5	9	31	1	2

The measured computational time requirements of the main steps for the different approaches are listed in Table III. Although the training of the SP+DTW approach is significantly quicker than the other references, the recognition part is slower due to running DTW comparison between the probe sample and all stored gait print samples. The LGEI approach needs relatively significant time for training set generation and training of the CNN and MLP networks, however the recognition step is still very efficient: less then 0.01sec/probe sample.

TABLE V
PRECISION/RECALL RATES OF ACTION RECOGNITION FOR EACH EVENT.

	Bend	Watch	Phone	Wave	Wave2
Sample num.	88	53	50	58	46
Precision	1.00	0.82	0.69	0.76	0.70
Recall	0.97	0.77	0.88	0.90	0.97

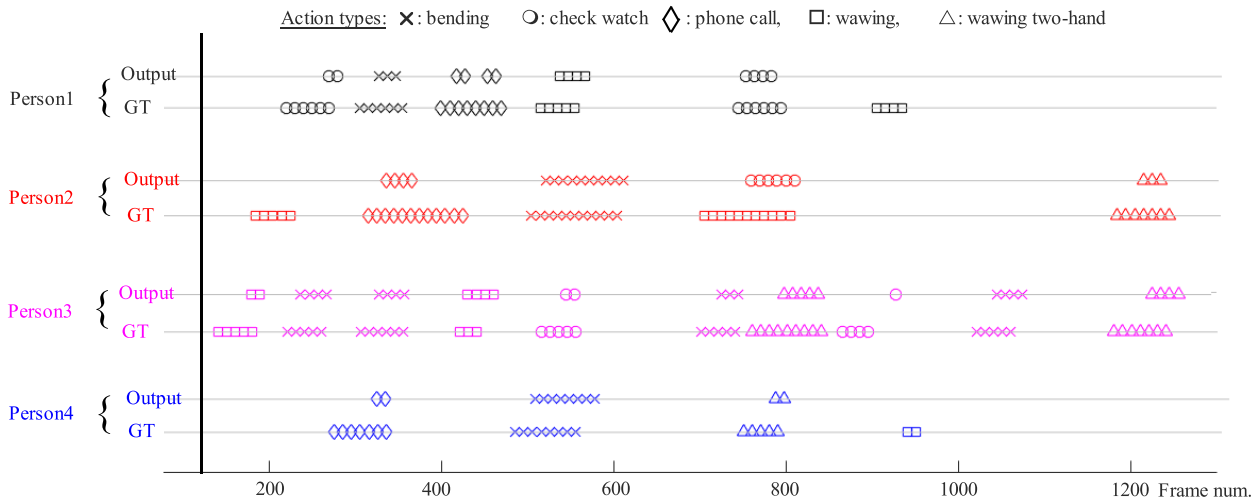


Fig. 17. Result of *activity recognition* in a selected outdoor test sequence (4 people). *Output* row: frames where our approach detected various activities, *GT* (Ground Truth): manually annotated reference frames. Each mark corresponds to 10 consecutive frames.

B. Evaluation of activity recognition

For evaluating the proposed activity recognition module, we used the ten activity sequences of the database, applying a cross validation approach. For testing the recognition performance on each sequence, we trained the actual CNNs with the manually annotated activity patterns of the other nine sequences. For both training and recognition we also used various negative samples cut from normal walking parts of the scenarios. The number of selected *walk* frames was equal to the average number of frames corresponding to the other activities.

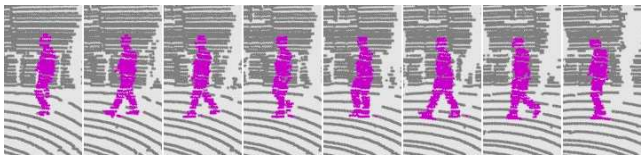
As the result, the aggregated confusion matrix of action recognition in the test scenes is shown in Table IV. The matrix value of the ith row and jth column indicates the number of samples from the ith activity, which were recognized as action j. The last two columns correspond to false negative (FN) and false positive (FP) detections, defined as follows for row i:

- FN: number of ignored occurrences of the ith action, which were neither identified by any of the other activities
- FP: number of erroneous alerts of the ith activity in the case when none of the addressed events occurred

As we can see, the *bend*, *phone*, *wave* and two-handed waving (*wave2*) activities were almost always denoted as an event (FN≤ 5), while check *watch* indicated 11 false negative samples, since the small hand movements were occasionally imperceptible due to occlusions. *Bend* was never confused with other actions, while *wave* and *wave2* were mixed up in a number of cases. It is also worth noting that the overall number of false positives is quite low (Σ_iFP$<5\%$ of the real events), i.e. the system rarely indicates unexpected warnings in case of normal walks. This advantageous property can be well examined in the timeline diagram displayed in Fig. 17, which corresponds to one of the outdoor test sequences. The horizontal axis corresponds to the frame index, and the different activities are denoted by different markers (as explained in the top row). For each person, the *Output* row marks the frames where our approach detected various activities, while the *GT* (Ground Truth) row indicate the manually annotated reference frames. In agreement with Table IV, in nearly all cases the real activities are detected by the system with a time delay necessary for ADM and AXOR map generation. Finally, Table V shows the one-vs-all detection precision and recall rates of each event separately, these cumulative results also confirm our above discussed experiences.

(a) Point cloud sequence (used for recognition)

(b) Video images sequence (not used, only visual reference)



(c) Synthetic 4D walk, gait phases synchronized with the Lidar observation

Fig. 18. Sample consecutive frames from the recorded (a) Lidar and (b) video sequences, and the synthesized 4D scene with leg movements synchronized with the observation

C. Demonstration of the visualized 4D scenario

In the visualization module of the 4D surveillance system the synchronization of the measurements and the steps of the

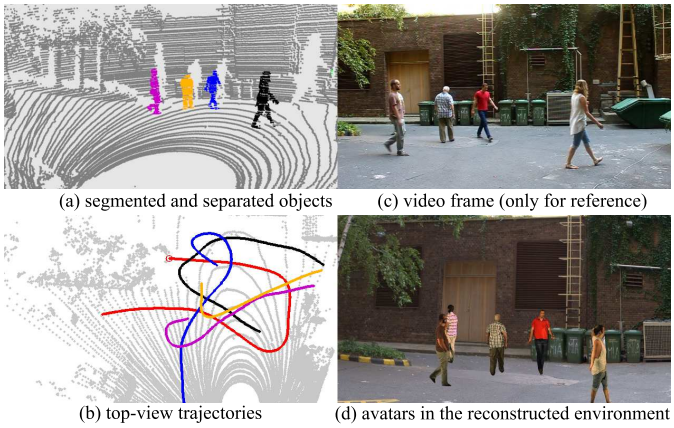


Fig. 19. Demonstration of the dynamic scene reconstruction process, with gait step synchronization.

animated avatars has been implemented. A sample sequence part is displayed in Fig. 18, showing simultaneously the processed Lidar point clouds, the reference optical video frames and the animated 4D studio object with synchronized steps with the observation. A summarizing figure of the complete recognition and visualization process is displayed in Fig. 19.

VIII. Conclusion

In this paper, we proposed algorithms for gait analysis and activity recognition based on the measurements of a RMB Lidar sensor, in realistic outdoor environments with the presence of multiple and partially occluded walking pedestrians. We provided quantitative evaluation results in various different measurement sequences, and demonstrated the advantages of the approach in possible future 4D surveillance systems. Future work will concern different types of RMB Lidar sensors, and fusion of various optical and 3D data sources in the recognition pipeline, such as Lidars or thermal cameras. The authors thank Levente Kovács from MTA SZTAKI for linguistic review and advices in deep learning issues.

References

[1] P.M. Roth, V. Settgast, P. Widhalm, M. Lancelle, J. Birchbauer, N. Brandle, S. Havemann, and H. Bischof, "Next-generation 3D visualization for visual surveillance," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2011, pp. 343 –348.

[2] D. Baltieri, R. Vezzani, R. Cucchiara, Á. Utasi, C. Benedek, and T. Szirányi, "Multi-view people surveillance using 3D information," in *Proc. International Workshop on Visual Surveillance at ICCV*, Barcelona, Spain, November 2011, pp. 1817–1824.

[3] L. Havasi, Z. Szlávik, and T. Szirányi, "Detection of gait characteristics for scene registration in video surveillance system," *IEEE Trans. Image Processing*, vol. 16, no. 2, pp. 503–510, Feb 2007.

[4] Z. Zhang, M. Hu, and Y. Wang, "A survey of advances in biometric gait recognition," in *Biometric Recognition*, vol. 7098 of *Lecture Notes in Computer Science*, pp. 150–158. Springer Berlin Heidelberg, 2011.

[5] W. Jin, M. She, S. Nahavandi, and A Kouzani, "A review of vision-based gait recognition methods for human identification," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Sydney, Australia, Dec 2010, pp. 320–327.

[6] M. P. Murray, "Gait as a total pattern of movement," *American Journal of Physical Medicine*, vol. 46, no. 1, pp. 290–333, 1967.

[7] H. Kim, J.-Y. Guillemaut, T. Takai, M. Sarim, and A. Hilton, "Outdoor dynamic 3-d scene reconstruction," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 11, pp. 1611 –1622, nov. 2012.

[8] C. Benedek, Z. Jankó, C. Horváth, D. Molnár, D. Chetverikov, and T. Szirányi, "An integrated 4D vision and visualisation system," in *International Conference on Computer Vision Systems (ICVS)*, vol. 7963 of *Lecture Notes in Computer Science*, pp. 21–30. Springer, St. Petersburg, Russia, July 2013.

[9] J. Hapák, Z. Jankó, and D. Chetverikov, "Real-time 4D reconstruction of human motion," in *Proc. 7th International Conference on Articulated Motion and Deformable Objects (AMDO 2012)*, 2012, vol. 7378 of *Springer LNCS*, pp. 250–259.

[10] C. Benedek, "3D people surveillance on range data sequences of a rotating Lidar," *Pattern Recognition Letters*, vol. 50, pp. 149–158, 2014, Special Issue on Depth Image Analysis.

[11] Y. Li, Y. Yin, L. Liu, S. Pang, and Q. Yu, "Semi-supervised gait recognition based on self-training," in *International Conf. Advanced Video and Signal-Based Surveillance (AVSS)*, Beijing, China, Sept 2012, pp. 288–293.

[12] Y. Makihara, H. Mannami, A. Tsuji, M.A. Hossain, K. Sugiura, A. Mori, and Y. Yagi, "The OU-ISIR gait database comprising the treadmill dataset," *IPSJ Trans. on Computer Vision and Applications*, vol. 4, pp. 53–62, Apr. 2012.

[13] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Cross-view and multi-view gait recognitions based on view transformation model using multi-layer perceptron," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 882 – 889, 2012.

[14] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, Feb 2006.

[15] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *International Conference on Computer Vision Workshops*, Kyoto, Japan, Sept 2009, pp. 1058–1064.

[16] D. Xu, S. Yan, D. Tao, S. Lin, and H. J. Zhang, "Marginal Fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2811–2821, 2007.

[17] Y. Huang, D. Xu, and F. Nie, "Patch distribution compatible semisupervised dimension reduction for face and human gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 3, pp. 479–488, 2012.

[18] D. Xu, Y. Huang, Z. Zeng, and X. Xu, "Human gait recognition using patch distribution feature and locality-constrained group sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 316–326, Jan 2012.

[19] Y. Huang, D. Xu, and T. J. Cham, "Face and human gait recognition using image-to-class distance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 3, pp. 431–438, 2010.

[20] A. Kale, N. Cuntoor, B. Yegnanarayana, A.N. Rajagopalan, and R. Chellappa, "Gait analysis for human identification," in *Audio- and Video-Based Biometric Person Authentication*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 706–714. Springer, 2003.

[21] Z. Liu and S. Sarkar, "Simplest representation yet for gait recognition: averaged silhouette," in *International Conference on Pattern Recognition*, Cambridge, UK, Aug 2004, vol. 4, pp. 211–214 Vol.4.

[22] W. Kusakunniran, Q. Wu, J. Zhang, Y. Ma, and H. Li, "A new view-invariant feature for cross-view gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1642–1653, 2013.

[23] H. Hu, "Multiview gait recognition based on patch distribution features and uncorrelated multilinear sparse local discriminant canonical correlation analysis," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 617–630, April 2014.

[24] S. Shirke, S.S. Pawar, and K. Shah, "Literature review: Model free human gait recognition," in *International Conf. Commun. Syst. and Network Technologies (CSNT)*, Bhopal, India, April 2014, pp. 891–895.

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

[26] S. Sarkar, P.J. Phillips, Z. Liu, IR. Vega, P. Grother, and K.W. Bowyer, "The humanID gait challenge problem: data sets, performance, and analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, Feb 2005.

[27] R. Gross and J. Shi, "The CMU Motion of Body (MoBo) Database," Tech. Rep. CMU-RI-TR-01-18, Robotics Instit., Pittsburgh, PA, 2001.

[28] Amit K. Roy Chowdhury, *Human Identification at a Distance – UMD Database*, 2001, available at `http://www.umiacs.umd.edu/labs/pirl/hid/data.html`.

[29] S. Sharma, A Shukla, R. Tiwari, and V. Singh, "View variations effect in gait recognition and performance improvement using fusion," in *IEEE*

International Conf. Recent Advances in Information Technology (RAIT), Dhanbad, India, March 2012, pp. 892–896.

[30] J. Tang, J. Luo, T. Tjahjadi, and Y. Gao, "2.5D multi-view gait recognition based on point cloud registration," *Sensors*, vol. 14, no. 4, pp. 6124–6143, 2014.

[31] T. Whytock, A. Belyaev, and N.M. Robertson, "Dynamic distance-based shape features for gait recognition," *Journal of Mathematical Imaging and Vision*, pp. 1–13, 2014.

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

[33] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits," *J. Vis. Comun. Image Represent.*, vol. 25, no. 1, pp. 195–206, Jan. 2014.

[34] T. Pallejá, M. Teixidó, M. Tresanchez, and J. Palacín, "Measuring gait using a ground laser range sensor," *Sensors*, vol. 9, no. 11, pp. 9133–9146, 2009.

[35] M. Teixidó, P. Mercé, T. Pallejá, M. Tresanchez, M. Nogués, and J. Palacín, "Measuring oscillating walking paths with a LIDAR," *Sensors*, vol. 11, no. 5, pp. 5071–5086, 2011.

[36] J. Ryu and S. Kamata, "Front view gait recognition using spherical space model with human point clouds," in *IEEE International Conf. Image Processing (ICIP)*, Brussels, Belgium, Sept 2011, pp. 3209–3212.

[37] L. Spinello, M. Luber, and K.O. Arras, "Tracking people in 3D using a bottom-up top-down detector," in *IEEE Int'l Conf. on Robotics and Automation (ICRA)*, Shanghai, China, 2011, pp. 1304–1310.

[38] C. Benedek, B. Nagy, B. Gálai, and Z. Jankó, "Lidar-based gait analysis in people tracking and 4D visualization," in *European Signal Processing Conference (EUSIPCO)*, Nice, France, September 2015, pp. 1143–1147.

[39] B. Gálai and C. Benedek, "Feature selection for Lidar-based gait recognition," in *Int'l Workshop on Computational Intelligence for Multimedia Understanding*, Prague, Czech Republic, October 2015.

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

[41] B. Kwolek, T. Krzeszowski, A. Michalczuk, and H. Josinski, "3D gait recognition using spatio-temporal motion descriptors," in *Intelligent Information and Database Systems*, vol. 8398 of *Lecture Notes in Computer Science*, pp. 595–604. Springer International Publishing, 2014.

[42] R.F. Mansour, "A robust approach to multiple views gait recognition based on motion contours analysis," in *National Workshop on Information Assurance Research (WIAR)*, April 2012, pp. 1–7.

[43] C. Benedek, D. Molnár, and T Szirányi, "A dynamic MRF model for foreground detection on range data sequences of rotating multi-beam Lidar," in *Advances in Depth Image Analysis and Applications*, vol. 7854 of *Lecture Notes in Computer Science*, pp. 87–96. 2013.

[44] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 22, pp. 747–757, 2000.

[45] D. Zhang and G. Lu, "A comparative study of Fourier descriptors for shape representation and retrieval," in *Asian Conference on Computer Vision (ACCV)*. 2002, pp. 646–651, Springer.

[46] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *International Joint Conference on Neural Networks (IJCNN)*, July 2011, pp. 1918–1921.

[47] A.W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, and M.F. Campos, "STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, vol. 7441 of *Lecture Notes in Computer Science*, pp. 252–259. Springer, 2012.

[48] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural svm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1422–1431, Aug 2013.

[49] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition," in *European Conf. Computer Vision (ECCV)*, vol. 8690 of *Lecture Notes in Computer Science*, pp. 742–757. Springer, 2014.

[50] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Portland, Oregon, June 2013, pp. 716–723.

[51] M. Munaro, G. Ballin, S. Michieletto, and E. Menegatti, "3D flow estimation for human action recognition from colored point clouds," *Biologically Inspired Cognitive Architectures*, vol. 5, pp. 42 – 51, 2013.

[52] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Mller, E. Säckinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," in *International Conference on Artificial Neural Networks*, Perth, Australia, 1995, pp. 53–60.

Csaba Benedek received the M.Sc. degree in computer sciences in 2004 from the Budapest University of Technology and Economics (BME), and the Ph.D. degree in image processing in 2008 from the Péter Pázmány Catholic University, Budapest. Between 2008 and 2009 he worked for 12 months as a postdoctoral researcher with the Ariana Project Team at INRIA Sophia-Antipolis, France. He is currently a senior research fellow with the Distributed Events Analysis Research Laboratory, at the Institute for Computer Science and Control of the Hungarian Academy of Sciences (MTA SZTAKI) and an associate professor with the Péter Pázmány Catholic University. He has been the manager of various national and international research projects in the recent years. His research interests include Bayesian image and point cloud segmentation, object extraction, change detection, and biometric identification.

Bence Gálai received the B.Sc. degree in 2015 from the Budapest University of Technology and Economics, where he is currently a M.Sc. student. He is also a junior researcher at the Distributed Events Analysis Research Laboratory of MTA SZTAKI since 2015. His research interest include biometric gait recognition in the wild, 3D video surveillance and point cloud registration. He is the co-creator of the first public Lidar based gait and activity database.

Balázs Nagy received the M.Sc. degree in computer engineering in 2016 from the Péter Pázmány Catholic University. He is a junior researcher and software developer at the Distributed Events Analysis Research Laboratory of MTA SZTAKI since 2013. He was the winner of a national B.Sc. thesis competition in 2014, and a key developer of the new semi-automatic traffic sign detection technology proposed for the Budapest Road Management Department.

Zsolt Jankó received the M.Sc. degree in Computer Science and Mathematics in 2001, and the Ph.D. degree in 2007, both from the Eötvös Loránd University, Budapest. In 2008 and 2009, he spent two years as postdoctoral researcher in France, in the CERTIS team at the Université Paris-Est and in the PERCEPTION group at INRIA Rhône-Alpes in Grenoble. He is currently a senior research fellow with the Distributed Events Analysis Research Laboratory at MTA SZTAKI. His main research interest focuses on computer vision and geometric modeling, including dynamic 3D object reconstruction, point cloud filtering and recognition, and large point cloud management. He has been the co-creator of the MTA SZTAKI 4D Reconstruction Studio.