An Integrated 4D Vision and Visualisation System

Csaba Benedek, Zsolt Jankó, Csaba Horváth, Dömötör Molnár, Dmitry Chetverikov and Tamás Szirányi*

Institute for Computer Science and Control, Hungarian Academy of Sciences
H-1111 Budapest, Kende utca 13-17, Hungary
firstname.lastname@sztaki.mta.hu

Abstract. This paper reports on a pilot system for reconstruction and visualisation of complex spatio-temporal scenes by integrating two different types of data: outdoor 4D data measured by a rotating multi-beam LIDAR sensor, and 4D models of moving actors obtained in a 4D studio. A typical scenario is an outdoor scene with multiple walking pedestrians. The LIDAR monitors the scene from a fixed position and provides a dynamic point cloud. This information is processed to build a 3D model of the environment and detect and track the pedestrians. Each of them is represented by a point cluster and a trajectory. A moving cluster is then substituted by a detailed 4D model created in the studio. The output is a geometrically reconstructed and textured scene with avatars that follow in real time the trajectories of the pedestrians.

Keywords: rotating multi-beam LIDAR, MRF, motion segmentation, 4D reconstruction

1 Introduction

Efforts on real time reconstruction of 3D dynamic scenes receive great interest in intelligent surveillance [13], video communication and augmented reality systems. 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 analysed from an arbitrary viewpoint, and virtually modified by the user. However, building an interactive 4D video system is highly challenging, as it needs in parallel automatic perception, interpretation, and real time visualisation of the environment.

A 4D reconstruction studio is an advanced, intelligent sensory environment, which uses multiple synchronized and calibrated high-resolution video cameras and a GPU to build dynamic 3D models providing free-viewpoint video in real-time. An example for this environment is introduced in [7]. While this system can efficiently record and visualise the model of a single moving person, in itself it is not appropriate to capture a large scenario with several moving people and various background objects.

* This work is connected to the i4D project funded by the internal R&D grant of MTA SZTAKI. Csaba Benedek also acknowledges the support of the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and the Grant #101598 of the Hungarian Research Fund (OTKA). Dmitry Chetverikov is also affiliated with Eötvös Loránd Univeristy (ELTE).

Recently a portable stereo system has been introduced [8] for capturing and 3D reconstruction of dynamic outdoor scenes. Here the observed scenario should be surrounded by several (8-9) carefully calibrated cameras beforehand, and the reconstruction process is extremely computation-intensive, as dealing with a short 10 sec sketch takes several hours. In addition, full automation is difficult due to usual stereo artefacts such as featureless regions and occlusions, which can cause significant problems in an uncontrolled outdoor environment.

Time-of-Flight (ToF) technologies, such us LIDAR, offer notable advantages versus conventional video flows for automated scene analysis, since in the provided 2.5D range data sequences geometrical information is directly available, and the measurements are significantly less sensitive on the weather and illumination conditions of the acquisition. High speed Rotating Multi-Beam (RMB) LIDAR systems, such as the Velodyne HDL-64E sensor, are able to provide accurate 3D point cloud sequences with a 15 Hz refreshing frequency, making the configuration highly appropriate for analysing moving objects in large outdoor environments with a diameter up to 100 meters. However, a single RMB LIDAR scan is a notably sparse point cloud, moreover we can also observe a significant drop in the sampling density at larger distances from the sensor and we also can see a ring pattern with points in the same ring much closer to each other than points in different rings [1]. These properties yield poor visual experiences for the observes, when a raw (Fig. 8(a)) or a semantically coloured (Fig. 8(c)) point cloud sequence is displayed in a screen.

The above observations motivated us to develop an unconventional system, called the *integrated 4D* (i4D) system, which combines two very different sources of spatio-temporal information, namely, a RMB-LIDAR and a 4D reconstruction studio. The main purpose of the integration of the two types of data is our desire to measure and represent the visual world at different levels of detail. In our approach, the LIDAR sensor provides a global description of a dynamic outdoor scene in the form of a time-varying 3D point cloud. The latter is used to separate moving objects from static environment and obtain a 3D model of the environment. The 4D studio builds a detailed dynamic model of an actor (typically, a person) moving in the studio. By integrating the two sources of data, which is to our best knowledge a unique attempt up to now, one can modify the model of the scene and populate it with the avatars created in the studio.

2 System description

This paper introduces the proposed i4D system, describing all major processing steps from the acquisition of the raw data (point clouds and videos) to the creation and visualisation of an augmented spatio-temporal model of the scene. The system configuration consists of the following processing blocks:

- *Data acquisition:* LIDAR based environment scanning for point cloud sequence generation,
- *Data preprocessing:* foreground and background segmentation of the LIDAR sequence by a robust probabilistic approach (Sec. 3.1),
- *Motion analysis:* detection and tracking of moving pedestrians, generating motion trajectories (Sec. 3.2),

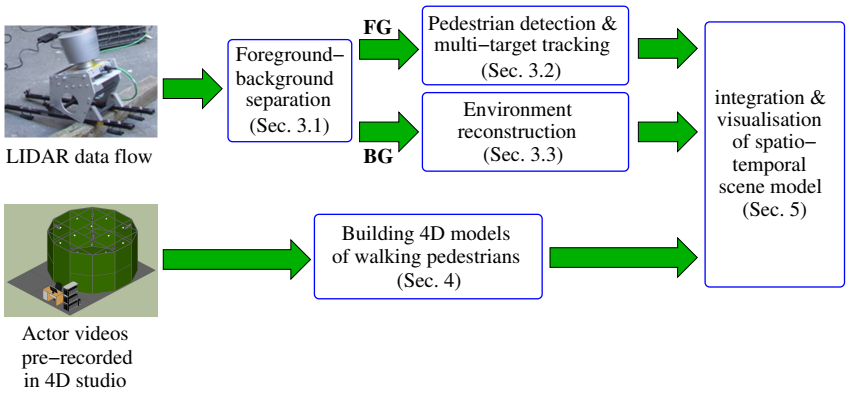

Fig. 1. Flowchart of the integrated 4D reconstruction system. **BG** is background, **FG** foreground.

- *Environment reconstruction:* geometric reconstruction of ground, walls and other field objects. Texturing the obtained 3D models with images of the scene (Sec. 3.3),
- *Pedestrian visualization:* creating textured moving pedestrian models in the 4D reconstruction studio (Sec. 4),
- *Integration:* transforming the system elements into a joint dynamic scene model and visualisation of the 4D scenario, where each avatar moves in the scene according to the assigned trajectory (Sec. 5).

Fig. 1 shows a flowchart of the complete i4D system. Each of the main building blocks is described in the corresponding section of the paper, as indicated in the flowchart.

3 LIDAR data processing

In this section, we present a hybrid method for dense foreground-background point labelling in a point cloud obtained by a Velodyne HDL-64E RMB-LIDAR device that monitors the scene from a fixed position. The method solves the computationally critical spatial filtering tasks applying an MRF model in the 2D range image domain. The ambiguities of the point-to-pixel mapping are handled by joint consideration of the true 3D positions and the 2D labels. Then, we execute detection and tracking of moving pedestrians for the foreground points. Next, we transform the background point cloud into a polygon mesh while maintaining the information about individual objects such as ground, walls, and trees. Finally, the models of the environment objects are manually textured using photos taken in the scene. Below, we describe these steps in more detail.

3.1 Foreground-background separation

The rotating multi-beam LIDAR device records 360°-view-angle range data sequences of irregular point clouds. Examples of measured point clouds will be shown later in this paper. To separate dynamic foreground from static background in a range data sequence, we apply a probabilistic approach [2].

range image foreground mask

Fig. 2. Example of foreground-background segmentation.

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. A part of a range image showing several pedestrians is demonstrated in Fig. 2. Spurious effects are caused by the quantisation error of the discretised view angle, the non-linear position corrections of sensor calibration, and the background flickering, e.g., due to vegetation motion.

One can model the dynamic range image as a Mixture of Gaussians and update the parameters similarly to the standard approach [14]. This provides a segmentation of the point cloud which is quite noisy because of the spurious effects. These effects are significantly decreased by the dynamic MRF model [2] that describes the background and foreground classes by both spatial and temporal features. The model is defined in the range image space. The 2D image segmentation is followed by a 3D point classification step to resolve the ambiguities of the 3D-2D mapping. Using a spatial foreground model, we remove a large part of the irrelevant background motion which is mainly caused by moving tree crowns. Fig. 2 shows an example of foreground segmentation.

3.2 Pedestrian detection and multi-target tracking

In this section, we present the pedestrian tracking module of the system. The input of the module step is a point cloud sequence, where each point is marked with a segmentation label of foreground or background. The output consists of clusters of foreground regions so that the points corresponding to the same person receive the same label over the sequence. We also generate a 2D foot point trajectory of each pedestrian to be used by the 4D scene reconstruction module.

First, the point cloud regions classified as foreground are clustered to obtain separate blobs for each moving person. We fit a regular lattice to the ground plane and project foreground regions onto this lattice. Morphological filters are applied in the image plane to obtain spatially connected blobs for different persons. Then we extract appropriately sized connected components that satisfy area constraints determined by lower and higher thresholds.

This procedure is illustrated in Fig. 3. The centre of each extracted blob is considered as a candidate for foot position in the ground plane. 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 artefacts at the level of the individual time frames, we developed a robust multi-tracking module which efficiently handles the problems at the sequence level.

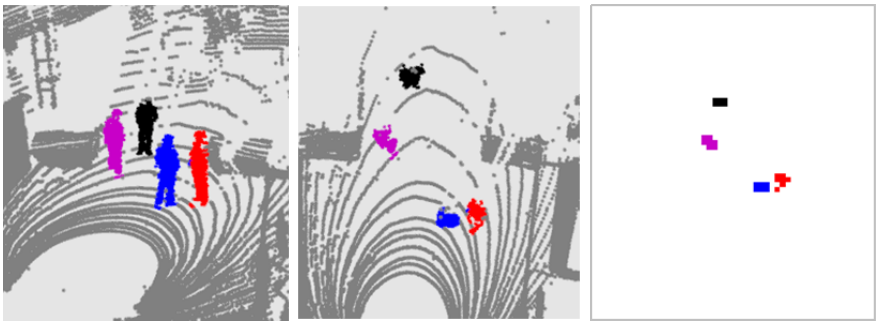

Fig. 3. Illustration of pedestrian separation.

Our multi-tracking algorithm receives the measured ground plane positions and for each frame iterates three basic operations, namely, data assignment, Kalman filter correction and Kalman filter prediction. The assignment operation assigns the candidate positions to objects, then the object positions are corrected and, finally, predictions for the subsequent positions are made and fed back to the assignment procedure. The algorithm can handle false positives as well as tracks starting and terminating within a sequence. Temporary track discontinuities are bridged in a post-processing step, while short false tracks are removed based on their length.

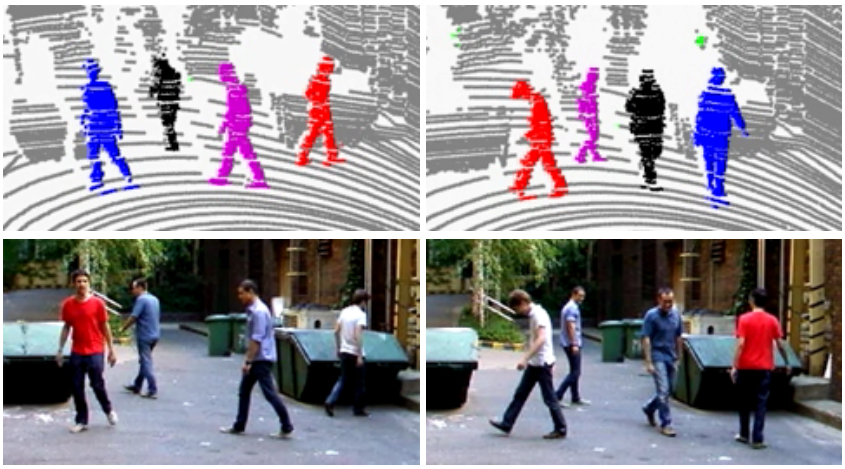

Fig. 4. Example of pedestrian tracking in a LIDAR sequence. Top row: point clusters whose colours identify the tracked persons. Bottom row: corresponding video frames displayed for verification.

The tracker module provides a set of pedestrian trajectories, which are 2D foot centre point sequences in the ground plane. To determine the points corresponding to each pedestrian in a selected frame, the connected foot blobs around a given trajectory point should be vertically back-projected to the 3D point cloud. A result of tracking is demonstrated in Fig. 4 that shows two segmented point cloud frames from a measurement sequence in a courtyard. It also shows the video frames taken in parallel as reference. One can observe that during the tracking the point cluster of a pedestrian preserves its colour.

3.3 Environment reconstruction

In this section, we describe our method for static environment reconstruction. First we accumulate the background points of the LIDAR sequence collected over several frames, which results in a dense point cloud that represents the ground, walls, trees, and other background objects. Assuming that the ground is reasonably flat and horizontal, we fit an optimal plane to this point cloud using the robust RANSAC [6] algorithm that treats all other objects as outliers. Points close to this plane are considered as ground points in the following. For vegetation detection and removal, we have developed an algorithm, which calculates a statistical feature for each point in the merged point cloud based on the distance and irregularity of its neighbors, and also exploits the intensity channel which is an additional indicator of vegetation, which reflects the laser beam with a lower intensity. The remaining points are then projected vertically to the obtained ground plane, where projections of wall points form straight lines that are extracted by the Hough transform [5]. Applying the Ball-Pivoting algorithm [3] to the 3D points that project to a straight line, we create a polygon mesh of a wall.

In the reconstruction phase, static background objects of the scene, such us trees, containers or parking cars are replaced with 3D models obtained from Google's 3D Warehouse. The recognition of these objects from the point cloud is currently done manually, and we are now working on the automation of this step. For example, one can adopt here the machine learning based approach of [10], which extracts various object level descriptors for point cloud blobs representing the detected objects, while to obtain similar representations of the training models from the 3D Warehouse, they perform ray casting on the models to generate point clouds, finally the classification is performed in the descriptor space.

Sample results of our environment reconstruction are shown in Fig. 8. Model texturing is based on a set of photographs taken in the scene.

4 Creating 4D models of walking pedestrians

Relatively small objects such as pedestrians cannot be reconstructed from the LIDAR range data in sufficient detail since the data is too sparse and, in addition, it only gives 2.5D information. Therefore we create properly detailed, textured dynamic models indoors, in a 4D reconstruction studio. The hardware and software components of such a studio can be found in [4, 7]. For completeness, we give below a brief description of the reconstruction process.

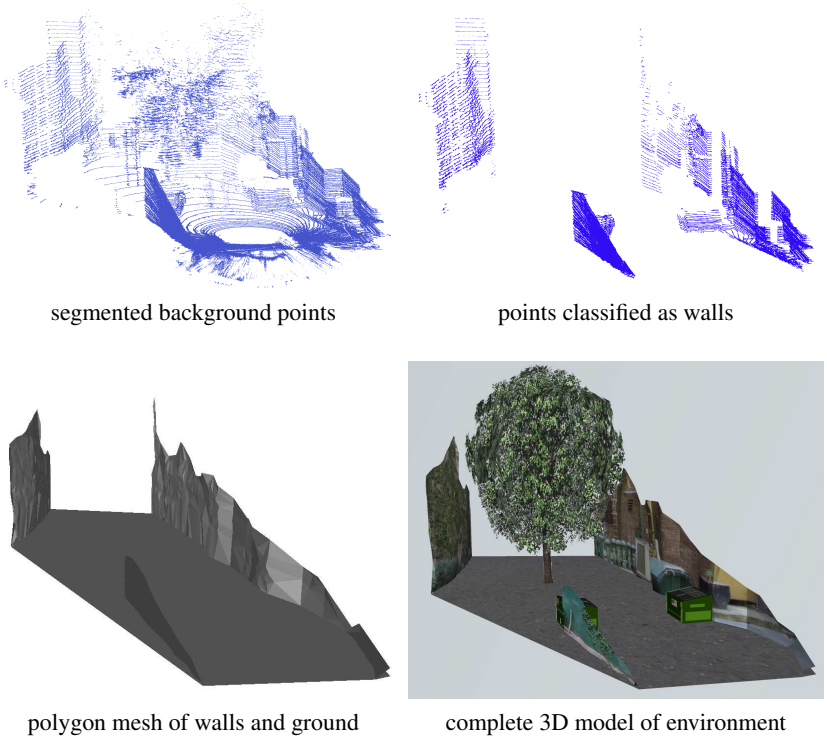

Fig. 5. Point cloud segmentation and environment reconstruction.

Fig. 6 shows a sketch and a panorama of the studio where green curtains and carpet form homogeneous background to facilitate segmentation of the actor. The frame carries 12 calibrated and synchronised video cameras placed uniformly around the scene, and one additional camera on the top in the middle. The cameras are surrounded by programmable LEDs that provide direct illumination. The studio has ambient illumination, as well. Seven PC-s provide the computing power and control the cameras and the lighting.

Currently, each set of 13 simultaneous video frames captured by the cameras is processed independently from the previous one. For a set of 13 images, the system creates a textured 3D model showing a phase of actor's motion. The main steps of the completely automatic 3D reconstruction process are as follows:

1. Colour images are extracted from the captured raw data.
2. Each colour image is segmented to foreground and background. The foreground is post-processed to remove shadows [4].
3. A volumetric model is created using the Visual Hull algorithm [11].
4. A triangulated mesh is obtained from the volumetric model using the Marching Cubes algorithm [12].
5. Texture is added to the triangulated mesh based on triangle visibility [7].

sketch panorama

Fig. 6. Sketch and panorama of a 4D reconstruction studio.

Fig. 7 shows an example of augmented reality created with the help of the 4D reconstruction studio. Several consecutive phases of an avatar walking in a virtual environment are displayed.

Fig. 7. A 4D studio actor walking in virtual environment.

5 Integrating and visualising the spatio-temporal scene model

The last step of the workflow is the integration of the system components and visualisation 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. Currently, we use the assumptions that the pedestrians walk forward along their trajectories. The top view orientation of a person is calculated from the variation of the 2D track.

To combine the 3D-4D data of different types arriving in different formats and visualise them in a unified format, we have developed a customised software system. All models are converted to the general-purpose OBJ format [15] which is supported by most 3D modelling programs and enables user to specify both geometry and texture.

Our visualisation program is based on the VTK Visualisation Kit [9]. Its primary goal is to efficiently support combining static and dynamic models allowing their multiplication and optimising the usage of computational resources. One can easily create mass scenes that can by viewed from arbitrary viewpoint, rotated and edited. Any user interaction with the models, such as shifting and scaling, is allowed and easy to perform.

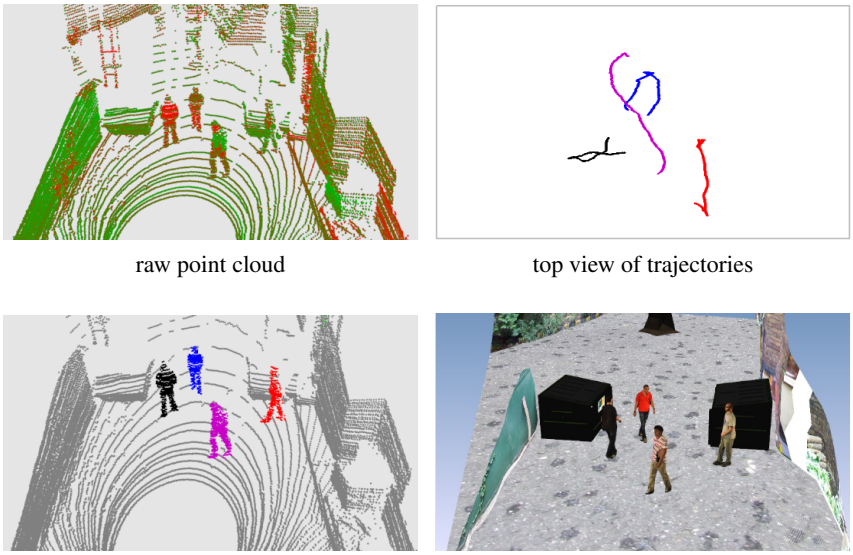

Fig. 8. Sample results of object tracking and integrated dynamic scene reconstruction.

The dynamic shapes can be multiplied not only in space, but in time, as well. Our 4D studio is relatively small. Typically, only two steps of a walking sequence can be recorded and reconstructed. This short sequence can be multiplied and seamlessly extended in time to create an impression of a walking person. To achieve this, the system helps the user by shifting the phases of motion in space and time while appropriately matching the sequence of the models.

An important requirement was to visualise the 3D motions of the avatars according to the trajectories provided by the LIDAR pedestrian tracking unit. An avatar follows the assigned 3D path, while rotation of the model to the left or right in the proper direction is automatically determined from the trajectory. Sample final results of the complete 4D reconstruction and visualisation process are demonstrated in Fig. 8.

6 Conclusion and outlook

In this paper, we have introduced a complex system on the interpretation and 4D visualisation of dynamic outdoor scenarios containing multiple walking pedestrians. As a key novelty, we have connected two different modalities of perception: a LIDAR point cloud stream from a large outdoor environment, and an indoor 4D reconstruction studio, which is able to provide detailed models of moving people. The proposed approach points towards real-time free-viewpoint and scalable visualisation of large scenes, which will be a crucial point in future augmented reality and multi modal communication applications. As future plans, we aim to extend the investigations to point

cloud sequences collected from a moving platform, and also implement automatic field object recognition and surface texturing modules.

References

1. Behley, J., Steinhage, V., Cremers, A.: Performance of histogram descriptors for the classification of 3D laser range data in urban environments. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 4391 –4398 (may 2012)
2. Benedek, C., Molnár, D., Szirányi, T.: A dynamic MRF model for foreground detection on range data sequences of rotating multi-beam lidar. In: International Workshop on Depth Image Analysis, LNCS. Tsukuba City, Japan (2012)
3. Bernardini, F., Mittleman, J., e.a.: The Ball-Pivoting algorithm for surface reconstruction. IEEE Transactions on Visualization and Computer Graphics 5(4), 349–359 (1999)
4. Blajovici, C., Chetverikov, D., Jankó, Z.: 4D studio for future internet: Improving foreground-background segmentation. In: IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 559–564. IEEE (2012)
5. Duda, R., Hart, P.: Use of the hough transformation to detect lines and curves in pictures. In: Comm. of the ACM. vol. 15, pp. 11–15 (1972)
6. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In: Comm. of the ACM. vol. 24, pp. 381–395 (1981)
7. Hapák, J., Jankó, Z., Chetverikov, D.: Real-time 4D reconstruction of human motion. In: Proc. 7th International Conference on Articulated Motion and Deformable Objects (AMDO 2012). Springer LNCS, vol. 7378, pp. 250–259 (2012)
8. Kim, H., Guillemaut, J.Y., Takai, T., Sarim, M., Hilton, A.: Outdoor dynamic 3-d scene reconstruction. IEEE Trans. on Circuits and Systems for Video Technology 22(11), 1611 –1622 (nov 2012)
9. Kitware: VTK Visualization Toolkit. `http://www.vtk.org` (2013)
10. Lai, K., Fox, D.: Object recognition in 3D point clouds using web data and domain adaptation. International Journal of Robotic Research 29(8), 1019–1037 (2010)
11. Laurentini, A.: The visual hull concept for silhouette-based image understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence 16, 150–162 (1994)
12. Lorensen, W., Cline, H.: Marching cubes: A high resolution 3D surface construction algorithm. In: Proc. ACM SIGGRAPH. vol. 21, pp. 163–169 (1987)
13. Roth, P., Settgast, V., Widhalm, P., Lancelle, M., Birchbauer, J., Brandle, N., Havemann, S., Bischof, H.: Next-generation 3D visualization for visual surveillance. In: IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS). pp. 343 –348 (30 2011-sept 2 2011)
14. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 747–757 (2000)
15. Wavefront Technologies: OBJ file format. Wikipedia, Wavefront .obj file (2013)