GMM-based Spatial Change Detection from Bimanual Tracking and Point Cloud Differences

Riccardo Monica, Andrea Zinelli, and Jacopo Aleotti

Robotics and Intelligent Machines Laboratory, Department of Information Engineering, University of Parma, Italy,
rmonica@ce.unipr.it
andrea.zinelli1@studenti.unipr.it
aleotti@ce.unipr.it

Abstract. Robots that detect changes in the environment can attain better context awareness and increased autonomy. In this work, a spatial change detection approach is presented which uses a single fixed depth camera to identify environment changes caused by human activities. The proposed method combines hand tracking and the difference between organized point clouds. Bimanual movements are recorded in real-time and encoded in Gaussian Mixture Models (GMMs). We show that GMMs enable change detection in presence of occlusions. We also show that the GMM analysis narrows down potential salient regions of space where manipulation actions are carried out. Experiments have been performed in an indoor environment for object placement, object removal and object repositioning tasks.

Keywords: Gaussian Mixture Models, range sensing, human motion tracking

1 Introduction

In this work, a method for spatial change detection is presented that identifies environment changes due to human activities. The experimental setup includes a single fixed depth camera (Kinect V2). Detection of the salient regions of the environment, where human actions have been carried out, is achieved by computing the difference between two organized point clouds, that are acquired at the beginning and at the end of each experimental session. Moreover, bimanual movements are tracked in real-time and encoded using GMMs. GMM analysis enables change detection in presence of occlusions and it reduces the number of false positives. Spatial change detection has mainly been investigated with object-based approaches by exploiting cameras or depth sensors [2, 9, 1, 6, 8, 13, 3, 7]. Petsch et al. [15] presented a framework for sensor-based detection of unexpected (surprising) events where manipulation events are detected from human observation and by placing markers on objects. In [4] GMMs were investigated for 3D data segmentation and novelty detection in the context of mobile robotics. Several authors proposed advanced approaches for segmentation of human hand

trajectories [11, 12, 5, 10, 16]. In particular, in [11] Gaussian Mixture Models have been applied for automatic segmentation of full-body motion trajectories.

2 GMM-based spatial change detection

2.1 Depth image processing

The proposed approach operates on two depth frames acquired automatically from a fixed Kinect V2, when no human motion is detected by the skeletal tracker: one (frame P) when the user has not yet entered the area, and the other (frame N) after the user left the scene. A difference depth image D_{uv} is computed (Eq. 1) given the one-to-one pixel correspondence in the two frames at coordinates (u, v).

$$D_{uv} = \begin{cases} \text{NaN} & \text{if } |N_{uv} - P_{uv}| \leq T_h \\ N_{uv} & \text{if } P_{uv} - N_{uv} > T_h \\ P_{uv} & \text{if } N_{uv} - P_{uv} > T_h \end{cases} \quad (1) \qquad I_{uv} = \begin{cases} \text{NaN} & \text{if } |N_{uv} - P_{uv}| > T_h \\ N_{uv} & \text{otherwise} \end{cases} \quad (2)$$

T_h is a threshold set from to the noise model of the sensor. According to Eq. 1, in case of a significant change between the depth values the nearest point is selected. Indeed, if the user performs an object placement task in direct sight of the camera depth image N contains relevant information about the newly placed object, while depth image P contains information about the background. The opposite occurs if an object is removed in direct sight of the camera. An invariant image is also computed as in Eq. 2, containing the pixels that do not change significantly. Both D_{uv} and I_{uv} are converted into point clouds called difference point cloud and invariant point cloud respectively.

2.2 Modeling of bimanual movements

Motion tracking of both hands is performed in real-time during the execution of the experiment using the Kinect V2 skeletal tracker. The trajectory of each hand is represented as a set of 3D points with time stamp, i.e. $\{x_k, y_k, z_k, t_k\}$. The proposed approach first generates two separate GMMs, M_L and M_R, one for each 4D hand trajectory using Expectation Maximization (EM). The number of Gaussians of M_L and M_R was chosen to minimize the BIC index [11, 14] as shown in Alg. 1. The two GMMs are then fused in a single GMM M. In particular, the priors w_i of the Gaussian components in M are computed as the original priors w_{Li} and w_{Ri} in M_L and M_R weighted by the ratio between the number of Gaussian components in M_L and M_R and the total number of Gaussians in M, i.e.: $w_i = w_{Li} \cdot |M_L|/|M|$ and $w_i = w_{Ri} \cdot |M_R|/|M|$. The weighting factor gives more importance to long trajectories. It turns out that isotropic (circularly symmetric) Gaussian components are likely to correspond to regions of space where salient manipulation activities are carried out. This fact can be explained by observing that user's actions are usually performed at slow speed and involve

Algorithm 1 Iterative Expectation Maximization

Input: S: set of points, with timestamp;
Output: $BestGMM$: the GMM model;
1: $MinBIC \leftarrow +\infty$; $GC \leftarrow 1$;
2: **repeat**
3: $GMM \leftarrow \text{EM}(S,GC)$;
4: $BIC \leftarrow \text{ComputeBIC}(GMM,S)$;
5: **if** $BIC < MinBIC$ **then**
6: $MinBIC \leftarrow BIC$;
7: $BestGMM \leftarrow GMM$;
8: **end if**
9: $GC \leftarrow GC + 1$;
10: **until** $BIC > MinBIC + BICTh$;

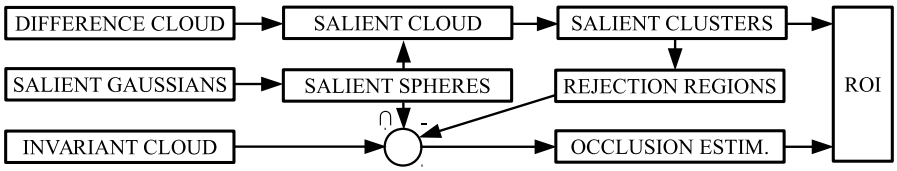

Fig. 1. Flowchart of the proposed approach for Regions of Interest (ROIs) extraction.

changes of hand direction, hence more points are sampled without a dominant direction. Thus, a Gaussian saliency value δ_i is computed as follows:

$$\delta_i = \frac{min_{j\in\{1,2,3\}}\sigma_{i,j}}{max_{j\in\{1,2,3\}}\sigma_{i,j}} \tag{3}$$

where $\sigma_{i,j}$ is the j-th eigenvalue of the covariance matrix. A Gaussian is considered salient if its prior w_i and its saliency δ_i are both greater than their average values in M, i.e.: $\left\{w_i > \bar{w} \wedge \delta_i > \bar{\delta}\right\}$.

2.3 Region of Interest extraction

Information extracted from depth image processing and GMMs is exploited to compute regions of interests (ROIs) corresponding to human activities as illustrated in Fig. 1. First, a set of salient spheres of fixed radius r is generated, each centered at the mean position of a salient Gaussian component in M. Then, a salient point cloud is computed as the part of the difference point cloud inside any of the salient spheres. Clusters of connected components in the salient point cloud (defined as salient clusters) are extracted. Salient clusters are likely to represent ROIs for manipulated objects in direct sight of the camera, i.e. objects that have been placed in the environment, moved or removed, as both human trajectory analysis and the difference point cloud agree. Each salient cluster generates a spherical ROI, centered at the cluster centroid with radius equal to the distance between the center and the farthest point of the cluster.

Salient spheres that contain at least T_I points of the invariant point cloud are also added to the list of the regions of interest. In fact, such spheres are likely to represent regions of space where user activity was detected from motion trajectory analysis although the Kinect depth frame could not locate any changes due to occlusions. Points of the invariant point cloud within a distance of r' from any salient cluster centroid (rejection region) are not counted to prevent duplicate ROI detections.

Fig. 2. Images of an example experiment (from left to right). The user picks up a red cone from inside a box (first salient action) and moves it on top of a table (second salient action). Images were recorded by an external camera close to the Kinect V2 sensor.

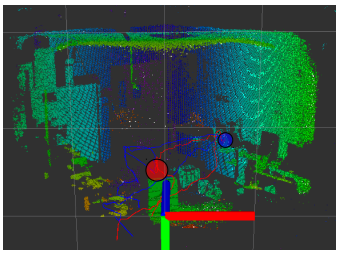

Fig. 3. Two regions of interest (blue and red sphere) are detected from the experiment in Fig. 2.

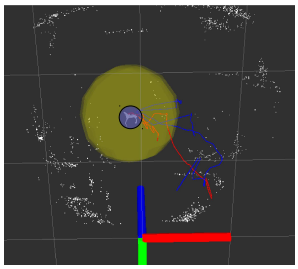

Fig. 4. A difference point cloud (white points) where sensor noise at the image border is filtered out by the GMM.

	In direct sight					Occluded				
Type	**TP**	**FP**	**FN**	**Prec**	**Rec**	**TP**	**FP**	**FN**	**Prec**	**Rec**
1	19	0	1	1.00	0.95	8	0	2	1.00	0.80
2	0	0	0	-	-	17	4	3	0.81	0.85
3	26	1	4	0.96	0.87	0	0	0	-	-

Table 1. True positives, false positives, false negatives, precision and recall for in direct sight and occluded actions.

3 Experiments

The proposed approach was evaluated by different users in an environment of size 5 by 4 meters. In each experiment the user entered the workspace, performed multiple manipulation actions and then left the scene. Fig. 2 shows an example experiment. Results of the spatial change detection algorithm are shown in Fig. 3. Fig. 4 illustrates, in a simpler experiment for clarity, the benefit of the GMM trajectory analysis when an action is performed in direct sight: the sensor noise at the depth image borders is filtered out. A quantitative evaluation was carried out on a dataset consisting of three experiments, with 10 trials for each experiment. Experiments of the first type consist of a sequence of three actions, two performed in direct sight of the sensor and one in an occluded region. The second type involves two relevant actions, both in occluded areas. Experiments of the third type consist again of three actions, all in direct sight of the sensor. Experiments have been performed with $r = 80\,cm$, $r' = 40\,cm$, $T_h = 10\,cm$ and $T_I = 50$. Results are summarized in table 3. Precision and recall are above 87% for actions in direct sight of the sensor and above 80% for actions in occluded regions.

References

1. E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter. Learning the semantics of object-action relations by observation. *Int. J. Rob. Res.*, 30(10):1229–1249, September 2011.
2. P. Alimi, D. Meger, and J.J. Little. Object persistence in 3D for home robots. In *The Semantic Perception, Mapping, and Exploration (SPME) workshop*, 2012.
3. R. Ambrus, N. Bore, J. Folkesson, and P. Jensfelt. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1854–1861, Sept 2014.
4. P. Drews, P. Núñez, R. Rocha, M. Campos, and J. Dias. Novelty detection and segmentation based on gaussian mixture models: A case study in 3d robotic laser mapping. *Robotics and Autonomous Systems*, 61(12):1696–1709, 2013.
5. D. R. Faria, R. Martins, J. Lobo, and J. Dias. Extracting data from human manipulation of objects towards improving autonomous robotic grasping. *Robotics and Autonomous Systems*, 60(3):396–410, 2012.
6. A. Fathi and J.M. Rehg. Modeling actions through state changes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
7. T. Fäulhammer, R. Ambrus, C. Burbridge, M. Zillich, J. Folkesson, N. Hawes, P. Jensfelt, and M. Vincze. Autonomous learning of object models on a mobile robot. *IEEE Robotics and Automation Letters*, 2(1):26–33, Jan 2017.
8. R. Finman, T. Whelan, M. Kaess, and J. J. Leonard. Toward lifelong object segmentation from change detection in dense RGB-D maps. In *2013 European Conference on Mobile Robots (ECMR)*, pages 178–185, Sept 2013.
9. E. Herbst, Xiaofeng Ren, and D. Fox. RGB-D object discovery via multi-scene analysis. In *IEEE/RSJ Intl Conference on Intelligent Robots and Systems (IROS)*, pages 4850–4856, Sept. 2011.
10. Sing Bing Kang and K. Ikeuchi. Toward Automatic Robot Instruction from Perception-Temporal Segmentation of Tasks from Human Hand Motion. *IEEE Transactions on Robotics and Automation*, 11(5):670–681, Oct 1995.
11. Sang Hyoung Lee, Il Hong Suh, S. Calinon, and R. Johansson. Learning Basis Skills by Autonomous Segmentation of Humanoid Motion Trajectories. In *IEEE-RAS Intl Conference on Humanoid Robots*, 2012.
12. J.F.-S. Lin and D. Kulic. Online Segmentation of Human Motion for Automated Rehabilitation Exercise Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):168–180, 2014.
13. J. Mason, B. Marthi, and R. Parr. Object disappearance for object discovery. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2836–2843, Oct 2012.
14. R. Monica, J. Aleotti, and S. Caselli. A kinfu based approach for robot spatial attention and view planning. *Robotics and Autonomous Systems*, 75, Part B:627–640, 2016.
15. S. Petsch and D. Burschka. Representation of manipulation-relevant object properties and actions for surprise-driven exploration. In *IEEE/RSJ Intl Conference on Intelligent Robots and Systems (IROS)*, 2011.
16. M. Yeasin and S. Chaudhuri. Toward automatic robot programming: learning human skill from visual data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 30(1):180–185, Feb 2000.