# 3D Point Cloud Segmentation Using a Fully Connected Conditional Random Field

Xiao Lin
Image Processing Group
Technical University of Catalonia (UPC)
Barcelona, Spain

Josep R.Casas
Image Processing Group
Technical University of Catalonia (UPC)
Barcelona, Spain

Montse Pardás
Image Processing Group
Technical University of Catalonia (UPC)
Barcelona, Spain

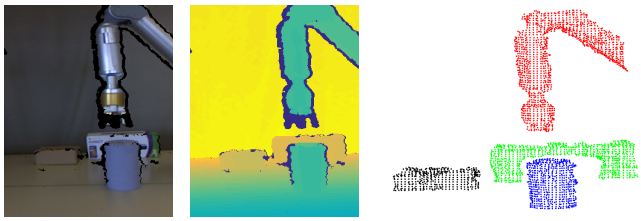
***Abstract*—Traditional image segmentation methods working with low level image features are usually difficult to adapt to higher level tasks, such as object recognition and scene understanding. Object segmentation emerges as a new challenge in this research field. It aims at obtaining more meaningful segments related to semantic objects in the scene by analyzing a combination of different information. 3D point cloud data obtained from consumer depth sensors has been exploited to tackle many computer vision problems due to its richer information about the geometry of 3D scenes compared to 2D images. Meanwhile, new challenges have also emerged as the depth information is usually noisy, sparse and unorganized. In this paper, we present a novel point cloud segmentation approach for segmenting interacting objects in a stream of point clouds by exploiting spatio-temporal coherence. We pose the problem as an energy minimization task in a fully connected conditional random field with the energy function defined based on both current and previous information. We compare different methods and prove the improved segmentation performance and robustness of the proposed approach in sequences with over 2k frames.**

## I. Introduction

Segmentation is an essential task in computer vision. It usually serves as the foundation for solving higher level problems such as object recognition, interaction analysis and scene understanding. Traditionally, segmentation is defined as a process of grouping homogeneous pixels into multiple segments on a single image, which is also known as low level segmentation. Low level approaches usually rely on features, such as color, texture, spatial relations between pixels on the image etc. to obtain segments which are somehow more homogeneous and more perceptually meaningful than raw pixels. Based on that, the concept of object segmentation is proposed. It is devoted to segment an image into regions which ideally correspond to meaningful objects in the scene. Contrarily to the low level segmentation tasks, object segments in the scene are not always related to a set of homogeneous pixels in the low level feature space, since an object may contain parts with very different appearance, be partially occluded by other objects or change its appearance with respect to the illumination. In this situation, more information is needed for tackling the problems in the object segmentation task.

### A. High level knowledge

High level knowledge is usually incorporated into the segmentation process to globally represent objects. For instance, object models are used in object segmentation for constrained scenes like human body detection [1]. However, most computer vision applications involve large amounts of data with different types of scenes containing several objects, which makes model based methods difficult to be adapted. In [2], accurate object annotation in the first frame is required to initially represent the objects in the scene for a video sequence. Propagating this representation along time provides a prior knowledge about the objects for the coming frames. More generally, some methods attempt to leverage the high level knowledge in more a generic fashion, such as [3], [4]. They train a classifier to determine whether an image region is an object-like region or not, which makes the approaches more generic to different types of objects and scenes. But these approaches strongly rely on the supporting of a large amount of labeled training data, which may not be available in some cases, such as with the relatively new consumer depth sensors.



(a) (b) (c)

Fig. 1. Point cloud segmentation in robotic arm grasping application. (a) Color image, pixels not related to valid depth value are removed. (b) Depth map, distance to the camera from close to far is marked with color from cold to warm. (c) Segmentation on a point cloud, different objects are marked with different color

### B. Temporal Information

The object segmentation task is usually tackled by employing temporal information when stream data is available, which is also known as video object segmentation. The temporal information serves as a hint for object segmentation in each frame, since pixels that belong to the same object are more likely to exhibit certain temporal coherence. For instance, Grundmann et al [5] use a graph-based model to hierarchically represent the spatio-temporal correspondences

ISBN 978-0-9928626-7-1 

between segments in the over-segmented frames and obtain a consistent video object segmentation by analyzing the hierarchy. Abramov et al [6] perform label transfer for pixels between frames by using optical flow. Then, they minimize the label distribution energy in the Potts model to generate labels for objects in the scene.

### C. Depth Information

Segmentation methods based on 2D images are limited, as a lot of valuable information about the shape and geometric layout of objects is lost when a 2D image is formed from the corresponding 3D world. The emergence of cheap consumer depth sensors, like Kinect, makes it easier to access depth information. This offers the potential to segment objects considering the richer geometric information in actual 3D. Problems such as occlusion and background removal are also easier to be tackled in 3D than in 2D images due to the extra dimension obtained from depth. Depth information provides the possibility to easily eliminate the background and focus on segmenting the interacting objects in the foreground, especially for object segmentation in applications such as vision based robotic arm grasping (shown in Fig.1) or semantic analysis in a smart room. However, 3D data also brings new challenges, because point clouds are usually noisy, sparse and unorganized. Furthermore, the lack of ground truth labeling for 3D point clouds also impairs the learning based approaches which require a large number of training data. In this situation, combining depth and temporal information emerges as a promising way to deal with the object segmentation task. Hickson et al [7] extend the hierarchical graph based method in [5] to RGBD stream data, in which depth maps are used to build another hierarchy similar to the color based one proposed in [5]. Abramov et al [8] obtained better results by introducing depth information in their label transfer approach [6]. In [9], we propose to model the 3D point cloud segmentation task as a labelling problem with a Conditional Random Field (CRF) model. We exploit temporal information by defining the energy function in the CRF model with respect to the object segmentation in the previous frame.

In this paper, we present a segmentation method for segmenting interacting foreground objects in applications like human manipulation in a smart room or robot arm grasping based on RGB-D stream data. In these applications we deal with a static camera targeting a specific space of interest which can be easily separated from the background using depth information. The aim is to segment the foreground area into time-consistent regions. This segmentation will allow to study the interactions of the objects in this area of interest. Our segmentation approach is initialized with the object mask in the first frame, which can be obtained automatically if the objects are spatially separated in the first frame. Apart from the first frame, we first extract the foreground region in the current frame, then we model the multi-objects foreground segmentation problem as an energy minimization task in a Fully Connected Conditional Random Field (FC-CRF) model. The energy function of FC-CRF is defined on both the object segmentation in the previous frame and current frame. It is optimized following an efficient inference method proposed in [10] to generate current object segmentation. We also compare different methods for achieving the time-consistent segmentation based on RGB-D stream data. We prove that the best results are obtained using Fully-Connected Conditional Random Field (FC-CRF) applied on super-voxels. We verify that super-voxels are more robustness than pixel based methods. We also show the improvement achieved using FC-CRF with respect to CRF.

## II. Graph based Objects Segmentation

The goal of multi-class image segmentation is to label every pixel in the image with one of several object categories. A common approach is to pose this problem as a graph segmentation problem. The graph is usually defined over pixels or image patches with the edges representing their spatial connectivity and the weight of the edges representing the similarity between them. Based on the graph representation, the minimum cut on the graph is searched to obtain several sub-graphs which are weakly connected to each other. These sub-graphs are assumed to be related to the objects in the image. Graph cut methods perform multi-class image segmentation without any prior information. However, when some prior information is available, the graph segmentation problem is usually posed as maximum a posteriori (MAP) inference in a Conditional Random Field (CRF)[11]. The CRF incorporates a pairwise energy term that maximizes label agreement between similar nodes on the graph, and a unary energy defined on each node representing the degree of belief that it belongs to each of the object categories with respect to the prior information. However, the pairwise energy is usually only computed for neighboring nodes on the graph in CRF, which makes the boundaries between different labels favor the thinner part of the graph with less edges. To overcome this limitation, Koltun et al [10] propose to employ Fully Connected CRF (FC-CRF) for image segmentation, in which all nodes on the graph are connected and the pairwise energy is established on any pair of nodes on the graph, which makes the shape of the graph less critical to the optimal labelling of the graph. They minimize the energy in FC-CRF using an efficient message passing implementation based on the mean fields approximation and high dimensional filtering.

## III. Data Acquisition and Foreground Extraction

Given the color and depth data captured by a consumer depth sensor, we can transform the per pixel distances provided in the depth image into a 3D point cloud $C_I \subseteq R^3$ with the camera parameters. As we are more concerned about the foreground cloud $C_{fg} \subseteq R^3$, we focus on the points in a Space of Interest (SoI). We define the SoI by manually setting a rectangular bounding box for each application. Points in the SoI are then treated as the foreground point cloud.

We need an initial object mask from which the time consistent segmentation of interacting objects will be built. This initial mask can be extracted automatically if the objects are

ISBN 978-0-9928626-7-1 © EURASIP 2017

separated in the initial frame by a simple connectivity analysis [12]. Otherwise, the initial mask has to be provided manually or by an external segmentation method.

## IV. Pixel Level Segmentation in RGB-D Stream Data

One possible approach for object segmentation in 3D point cloud is to extend the FC-CRF method for 2D images to 3D point clouds, since they are "1 to 1" related. We denote this extended method as Pixel Level Fully Connected Conditional Random Field (P-CF-CRF). In practice, the energy function in P-FC-CRF is modeled as the summation of a unary energy and a pairwise energy. We define the unary energy for each pixel (a point on the point cloud) with respect to the prior information, thus the object segmentation in the previous frame, and the pairwise energy based on similarity between pixels in color and location.

$$E(L) = \sum_{v_i \in v, l_i \in l} E_u(v_i, l_i) + \sum_{v_i, v_j \in v} E_p(v_i, v_j) \quad (1)$$

where $l_i \in L$ represents the label for a 3D point on the point cloud (a pixel) $v_i$.

We define the unary energy of labelling point $v_i$ with object label $o_j$, $E_u(v_i, l_i = o_j)$ proportional to the mean 3D Euclidean distance between point $v_i$ in the current point cloud and the $k$ nearest points labeled by $o_j$ in the previous point cloud. We compare the distance based unary energy with a more complicated unary energy term defined on multi-features (color, location and local shape) in the experiment section. It proves that FC-CRF allows to achieve comparable segmentation performance with simple energy terms like the distance based unary energy. For the pairwise energy, we extend the one defined in [10] for pixels on 2D image to an energy which is suitable for nodes representing 3D points. Specifically, we adopt an appearance and an smoothness term balanced with weights $\omega_1$ and $\omega_2$. The appearance energy term is defined as the 3D Euclidean distance $d_e(v_i, v_j)$ between points $v_i$ and $v_j$, and the color difference $d_{rgb}(v_i, v_j)$ between them in the Gaussian kernel $\exp\left(-\frac{d(\cdot)}{2\sigma^2}\right)$. The smoothness energy term is defined as the 3D Euclidean distance between the two points in the Gaussian kernel. That is:

$$\begin{aligned} E_p(v_i, v_j) =& \omega_1 \exp\left(-\frac{d_e(v_i, v_j)}{2\sigma_\alpha^2} - \frac{d_{rgb}(v_i, v_j)}{2\sigma_\beta^2}\right) \\ &+ \omega_2 \exp\left(-\frac{d_e(v_i, v_j)}{2\sigma_\gamma^2}\right) \end{aligned} \quad (2)$$

The appearance term favours nearby points in the 3D space to be in the same class if they have similar color. The smoothness term removes isolated regions, as in [10]. As shown in Eq.2, $\omega_1$ and $\omega_2$ are used to balance the appearance energy and smoothness energy. $\sigma_\alpha$, $\sigma_\beta$ and $\sigma_\gamma$ control the "scale" of the Gaussian kernel. In our experiments, $\omega_1$ and $\omega_2$ are set to $0.6$ and $0.4$ empirically. $\sigma_\alpha$ and $\sigma_\gamma$ are set to $0.3$ meter with respect to its physical meaning. We set $\sigma_\beta$ to 13 following [10]. The energy function in Eq.1 is minimized using the mean fields approximation and high dimensional filtering proposed in [10]. The optimum represents the best labelling on the graph, which also corresponds to the object segmentation of the point cloud.

(a) (b)

Fig. 2. An example of the super-voxels generated in our approach from a point cloud. (a) The original point cloud. (b) The super-voxels, each super-voxel is marked with a random color.

## V. Super Voxel Construction in RGB-D Data

In an image segmentation task, a mid-level representation is often employed to improve the robustness to noise in the raw data while keeping sufficient information of the image structure and flexibility for image segmentation. In particular, super-pixels is one of the most promising mid-level representations in image segmentation [13], [14]. These methods group homogeneous pixels on an image into super-pixels while preserving explicit boundary information of object parts. Similar to super-pixel methods, there are also several 3D approaches working on grouping 3D points represented by local features into meaningful 3D segments based on RGB-D data, such as super-voxels [15], or region growing method [16]. They usually preserve better object boundaries than similar 2D methods due to significative boundary information provided by depth maps. In [9], several mid-level representations are compared in a point cloud segmentation task. The super-voxel method proposed in [15] proved to be the best representation among the methods tested. It is based on first voxelizing the 3D points obtained from RGB-D data in the 3D space, then grouping the voxels into super-voxels with respect to their proximity in 3D space, color similarity and local 3D shape similarity. Thus, we employ the method in [15] to generate super-voxels in our approach. Fig.2 shows an example of super-voxels (labelled in random color) obtained from the original point cloud.

## VI. Super Voxel Level Segmentation with in RGB-D Stream Data

The obtained super-voxels are more perceptually meaningful than raw pixels. Performing object segmentation at the super-voxel level has a smaller problem scale and is more robust to noise in the raw data. In Sec. IV, we have proposed an extension of FC-CRF to segment point clouds. In this Section we propose to apply the FC-CRF at the super-voxel level, thus working with the complete super-voxel graph. The energy function here is defined in the same manner than in Sec.IV. The unary energy for each super-voxel is calculated with respect to the object segmentation in the previous frame, and the pairwise energy is computed as the similarity between

ISBN 978-0-9928626-7-1 © EURASIP 2017

| Seq. No. | nFrames | CRF | P-FC-CRF | S-FC-CRF |
|---|---|---|---|---|
| 1 | 601 | 94.65 | 94.34 | **97.93** |
| 2 | 425 | 96.02 | 95.07 | **97.42** |
| 3 | 747 | 95.72 | 93.30 | **97.15** |
| 4 | 291 | 94.72 | 94.04 | **96.50** |
| average | 516 | 95.27 | 94.19 | **97.25** |

TABLE I
MEAN IOUS IN 4 SEQUENCES OF RGBD VIDEOS PRODUCED BY METHOD IN [9], P-FC-CRF AND S-CF-CRF

super-voxels in color and location. Since a super-voxel represents a set of 3D points, we define the location for a super-voxel as the centroid of the related set of 3D points and the color as the mean color of the 3D points.

## VII. EXPERIMENTS

To evaluate the proposed 3D point cloud segmentation method, we employ 4 sequences in the Human Manipulation dataset [17] with the ground truth labelling provided in [9]. These 4 sequences focus on scenes where a human interacts with multiple foreground objects, and contain over 2k frames with challenges like occlusion, fast moving objects and multi-objects interaction. The ground truth provides object labels for points on the point cloud in all the sequences. The evaluation metrics is Intersection over Union (IOU) on point cloud, defined as follows:

$$IOU = \frac{1}{M_o} \sum_{m=1}^{M_o} \max_i \frac{GT_m \cap O_i}{GT_m \cup O_i} \tag{3}$$

For each frame, $M_o$ stands for the number of objects labeled in the ground truth, $GT_m$ is the ground truth for object $m$ and $O_i$ represents an object segmentation in this frame.

In a first experiment, we verify the performance of three point cloud segmentation methods, which are Conditional Random Field based method (CRF) proposed in [9], Pixel level (Sec.IV), and Super-voxel level (Sec.VI) Fully Connected Conditional Random Field based methods proposed in this paper (P-FC-CRF,S-FC-CRF respectively), with accurate prior information. For this initial verification we perform each frame segmentation based on the ground truth object labeling of the previous frame. Table I shows the segmentation performance of these three methods. S-FC-CRF achieves the best results in all sequences and obtains $2\%$ improvement in total with respect to CRF based method in [9].

In the second experiment, we analyze the segmentation error for the whole sequence when we initialize the system with ground truth object labelling *only for the first frame* in each of the 4 sequences. We analyze in Fig.3 the accumulated error in this situation. The P-FC-CRF method is not compared in this experiment due to its limited robustness to the accumulated segmentation error, which we justify in the following experiment. Fig.3 shows the segmentation result from CRF and S-FC-CRF (marked in red and blue respectively), in which we shows the mean IOU in the vertical axis over the frames up to frame $t$ in the horizontal axis. The curve represents the trend of the segmentation performance. Our S-FC-CRF based

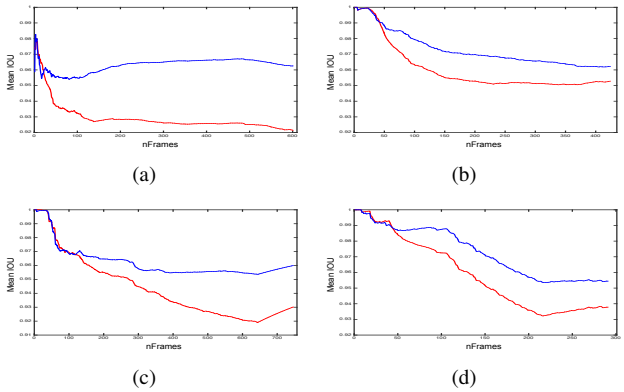



Fig. 3. Segmentation performance shown as mean IOU (vertical axis) over $n$ frames (horizontal axis) in 4 different sequences. Red: method in [9]. Blue: S-FC-CRF.

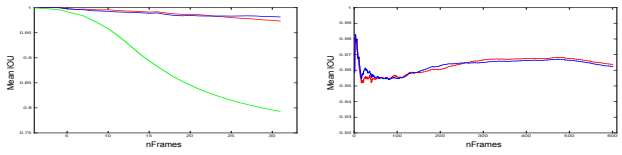



Fig. 4. Segmentation performance verification: (a) on robustness to segmentation error for P-FC-CRF in green compared to CRF and S-FC-CRF in red and blue, (b) on employing different unary energy in S-FC-CRF, shown as mean IOU (vertical axis) over $n$ frames (horizontal axis). Blue: S-FC-CRF with unary energy defined based only on difference in location. RED: S-FC-CRF with unary defined based on difference in color, local surface normal and location.

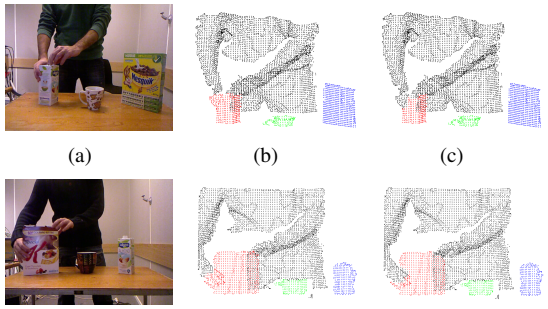
method keeps the segmentation performance at a higher level of mean IOU while also decaying slower than the CRF based method, which proves its stronger robustness with respect to the accumulated segmentation error.

In Fig.4(a), we show the performance of P-FC-CRF in the same manner than Fig.3. Compared to CRF and S-FC-CRF methods, it decays from 1 to 0.7 within around 30 frames, which also proves that the super-voxel representation employed in CRF and S-FC-CRF provides some robustness to the accumulated segmentation error.

As mentioned in Sec.IV, we evaluate the impact of employing different unary energies in FC-CRF. Specifically, we compare the unary energy defined only based on distance in our approach with a more complex unary energy which includes color, local surface normal and location in the S-FC-CRF method. Fig.4(b) illustrates that using the simple distance based unary energy achieves comparable segmentation performance ($0.2\%$ lower) than the more complex one.

Fig.5 presents some qualitative results of the CRF method and the proposed S-FC-CRF method. Each row in Fig.5 shows the point cloud segmentation in one frame. In Fig.5(b), the segmentation in CRF method favors the thinner part of the graph with less edges on the wrist, which leads the segmentation error on the palm. However our S-FC-CRF based method provides better segmentation on the boundary due to the fact

ISBN 978-0-9928626-7-1 



(a) (b) (c)

(d) (e) (f)

Fig. 5. Qualitative results. First column: color images, second column: results from [9], third column: results from S-FC-CRF

(a) (b) (c)

Fig. 6. Qualitative results from S-FC-CRF for robotic arm grasping.

that fully connected structure make the graph shape less critical for segmentation (shown in Fig.5(c)). A similar example is presented in the second row. We also show the potential of the proposed approach in a robotic arm grasping task. We applied the proposed method on a RGB-D sequence captured in the scene where a robotic arm moves to grasp objects on the table. Qualitative results are shown in Fig.6. More visual results are available on: https://imatge.upc.edu/web/node/1868.

Apart from the evaluation on segmentation performance, we compare the computation cost among CRF, S-FC-CRF and P-FC-CRF by calculating the average time cost per frame for segmenting 600 frames. Table II shows that CRF based method has the lowest average computation cost, however the computation cost in S-FC-CRF is also well handled by the efficient inference method proposed in [10], which achieves higher segmentation accuracy in similar time.

## VIII. Conclusion

In this paper, we have introduced a temporally consistent 3D point cloud segmentation method, in which we pose the segmentation task as energy minimization in a fully connected conditional random field. We define the unary energy term based on previous information for a super-voxel representation of a point cloud to impose temporal coherence while exploiting the pairwise energy established on the complete set of nodes to reduce the critical impact of the graph shape in segmentation.

| CRF | S-CF-CRF | P-CF-CRF |
|---|---|---|
| 0.015s | 0.021s | 0.07s |

TABLE II
Computation cost

The evaluation of the proposed approach is done quantitatively over 2k frames. Comparisons are made between the method in [9] and the proposed method, which illustrates the stronger robustness and better segmentation performance of the proposed approach. We achieve $2\%$ and $4\%$ improvement respectively in the first two experiments shown in Sec.VII with respect to method in [9]. We also show the potential of the proposed approach in an application of robotic arm grasping.

## References


[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[2] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label mrf optimization," *IJCV*, vol. 100, no. 2, pp. 190–202, 2012.

[3] H. Fu, D. Xu, and S. Lin, "Object-based multiple foreground segmentation in rgbd video." *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 2017.

[4] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[5] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2141–2148.

[6] A. Alexey, E. E. Aksoy, J. Dörr, F. Wörgötter, K. Pauwels, and B. Dellen, "3d semantic representation of actions from effcient stereo-image-sequence segmentation on gpus," in *International Symposium 3D Data Processing, Visualization and Transmission (3DPVT) Edition 5th*, 2010, pp. 1–8.

[7] S. Hickson, S. Birchfield, I. Essa, and H. Christensen, "Efficient hierarchical graph-based segmentation of rgbd videos," in *CVPR2014*. IEEE Computer Society, 2014.

[8] A. Abramov, K. Pauwels, J. Papon, F. Wörgötter, and B. Dellen, "Depth-supported real-time video segmentation with the kinect," in *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*. IEEE, 2012, pp. 457–464.

[9] X. Lin, J. Casas, and M. Pardàs, "3d point cloud segmentation oriented to the analysis of interactions," in *The 24th European Signal Processing Conference (EUSIPCO 2016)*. Eurasip, 2016.

[10] V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Adv. Neural Inf. Process. Syst*, vol. 2, no. 3, p. 4, 2011.

[11] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[12] X. Lin, J. R. Casas, and M. Pardàs, "3d point cloud video segmentation based on interaction analysis," in *Computer Vision–ECCV 2016 Workshops*. Springer, 2016, pp. 821–835.

[13] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[14] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *European Conference on Computer Vision*. Springer, 2008, pp. 705–718.

[15] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter, "Voxel cloud connectivity segmentation-supervoxels for point clouds," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2027–2034.

[16] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas, "Detecting end-effectors on 2.5 d data using geometric deformable models: Application to human pose estimation," *Computer Vision and Image Understanding*, vol. 117, no. 3, pp. 281–288, 2013.

[17] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellstrom, "Audio-visual classification and detection of human manipulation actions," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 3045–3052.


ISBN 978-0-9928626-7-1 