RISK ASSESSMENT FOR RGBD SCANS IN REAL TIME

R. Dupre, V. Argyriou, D. Greenhill

Kingston University

ABSTRACT

In this paper we address the notion of risk assessment of three dimensional scenes. Furthermore through the use of local feature recognition techniques and machine learning we perform this analysis on real time point cloud recordings. We provide a definition of risk and potential hazards that incorporates different elements but mainly focuses on intrinsic risk related properties of an object (e.g sharpness). A 3D Voxel HOG descriptor is utilised that aims to classify and recognise the presence of hazardous characteristics and features of objects present in a given scene. Additionally we utilise and extend the 3D Risk Scenes Dataset (3DRS) designed for risk evaluation in scene analysis. The effectiveness of our method is tested on captured point cloud sequences containing hazardous and non hazardous data with a high degree of accuracy across all tested data.

***Index Terms*—** Scene Analysis, 3DHOG, Real Time

1. INTRODUCTION

Scene analysis is a research area covering a large range of topics with applications in traffic analysis [1], domestic robotics [2], smart homes [3] and more recently in risk detection [4, 5], amongst many others. With the recent availability of depth sensors [6], the extension into 3D comes allows the inference of information about a scene not possible before.

In this work we consider the problem of evaluating risk in a three dimensional scene in real time and providing a quantified risk score. This translates into real world applications such as domestic robotics focusing on providing child care or supporting elder people by identifying hazardous situations. Up until now the research has focused on static 3D scenes, and often on fully realised 3D models of those scenes. This however does not effectively represent a realistic approach to the problem, as the intended applications for this research are unlikely to have access to that quality of data. As such within his work an attempt has been made to make an analysis on streamed data from depth hardware.

The definition of a risk or hazard in an environment is contextual. As an example consider a ball and knife placed on a table. The position of those objects relative to the edge means they may be easier to dislodge and therefore present more of a hazard. However the object itself can add to that risk; a knife will always have a blade (Figure 1).

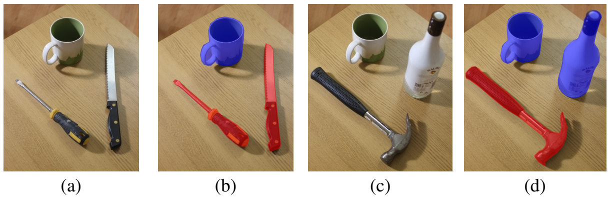


Fig. 1: Scenes of objects with intrinsic properties (e.g. sharp, pointed) and the goal identification of risky (red) objects versus safe (blue).

To identify these hazardous properties we utilise the 3D Voxel HOG feature descriptor based on the principles of Histogram of Oriented Gradients. With the proposed descriptor and boosting techniques (Adaboost [7]) we look to classify the objects as hazardous or not in a realtime situation. This classification can be used as a basis for a Risk Estimation Framework that could further define risk based on the wider context. Importantly object recognition is not the goal, allowing the proposed approach to be more general and operate at a lower level. In this work we define 'hazardous features' as any structure present in an object that could increase risk.

The paper will continue as follows; in section 2 an analysis of the similar areas of research and related work. An analysis of the proposed methodologies and processes used will be presented in section 3. Section 4 will outline the experiment environments and analyse the results. Finally, in section 5 conclusions are drawn.

2. RELATED WORK

2.1. Risk Assessment in Scenes and 3D Descriptors

To date very little research has been done in automated risk analysis systems. [8, 9] analyse indoor fall assessment for elderly adults; here both proposed methods focus on analysing the person themselves not the risk of the scene. [4] introduces the notion of analysing the fall potential of objects in a scene given the influence of environmental events such as human intervention or earthquakes. However the hazardous features of the objects themselves are not analysed.

Another emerging area of research within 3D scene analysis and understanding is Volumetric Reasoning. Here further information about a scene can be derived by applying certain logic based algorithms to object clusters in a given volume, leading to improvements in segmentation and clustering techniques. [10] applies the notion that clusters in a scene should

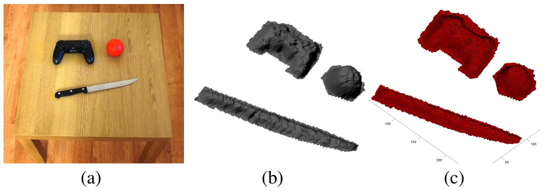

(a) (b) (c)

Fig. 2: (a) original scene, (b-e) result of acquisition process and planar removal, (f) voxelization process.

be at rest when simulation techniques are applied. Using an iterative process object clusters are grouped until such time as the scene is at equilibrium. [11] proposes a method that better fits bounding shapes to RGB-D clusters based on the premise that a good 3D representation of a scene is stable, fits the data well and is self-supporting. It is worth mentioning the following papers that consider similar concepts and approaches for scene analysis [12, 13, 14, 15]. Although the concept of risk in the environment is raised in some of this work, a real time automated form of risk evaluation is not addressed.

Object retrieval and feature selection are research subjects that have received a huge amount of work in recent years both in the 2D and 3D domains. The initial concept of HOG features [16] revolutionized the 2D object recognition world by creating a local descriptor that was resistant to geometric and photometric changes. Felzenszwalb [17] proposed a highly accurate object detection method through the use of deformable part models and HOG features. Buch in [18] implements a vehicle recognition framework using a patch definition system on a 3D representation of the found vehicle. This is combined with a traditional two dimensional HOG implementation.

With the introduction of financially viable 3D depth camera hardware, such as the Microsoft Kinect [6], more research has been focusing in the 3D domain. Transferring Dalal and Triggs work into three dimensions, Scherer [19] performs gradient computation in 3D using a convoluted distance field. This provides an effective way of calculating the magnitudes of the gradients, scoring them highly when localised near a surface of a model (local maxima). However their method also scores highly those at local minima creating additional artifacts within the data making it unsuitable for local feature recognition. Another example that uses a variation of vectors within a histogram as a feature is [20]. Here the normal vectors are used as the feature to define an object. Additionally the following state of the art 3D descriptors should be mentioned [21, 22].

3. PROPOSED METHODOLOGY

3.1. Pre-Processing

As analysis is to be in a real time fashion, a sequential set of frames capturing a scene is used. For each frame a 3D mesh model reconstruction of the scene is captured using Kinect 2 [23] (Figure 2b). Analysis is done on a frame by frame basis. As such before the risk in a frame can be evaluated, pre-processing steps are required to convert the captured mesh model of a frame into a usable format (Figure 2).

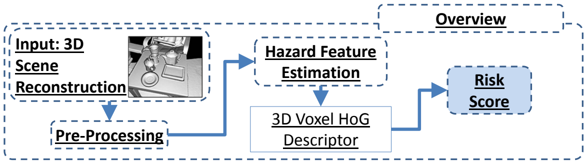


Fig. 3: An overview of the proposed risk assessment framework.

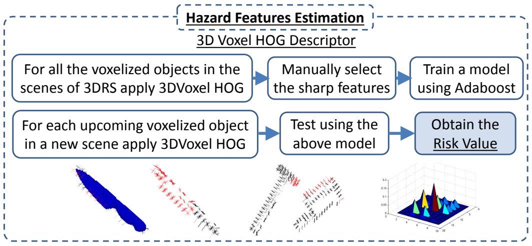

Hazard Features Estimation
3D Voxel HOG Descriptor
For all the voxelized objects in the scenes of 3DRS apply 3DVoxel HOG
Manually select the sharp features
Train a model using Adaboost
For each upcoming voxelized object in a new scene apply 3DVoxel HOG
Test using the above model
Obtain the Risk Value

Fig. 4: The proposed hazard classification system.

As the angle of the scene changes relative to the recording device from frame to frame the mesh models must be aligned to aid the analysis process [24]. Additionally the surface on which the objects are set requires removal, the work in [25] provides solutions for these problems. Voxelisation of the mesh model[26] is then performed. Voxels are defined along the faces of the scene's 3D mesh, voxels enclosed within a mesh are also classified as part of the object, allowing us to consider features based on an objects' density (Figure 2c). The resultant scene volume representing a binary classification of either object or not.

3.2. Introduction of a Risk Estimation Framework

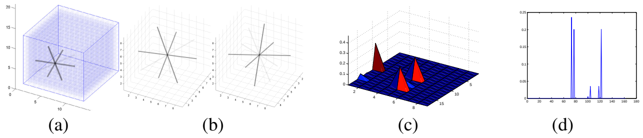
Let us define the cumulative risk score R for a scene as the weighted sum of n risk elements E as

$$R = \sum_{i=1}^{n} w_i E_i \quad (1)$$

A risk element is any measure that could highlight potential risk. These elements could include concepts such as stability [27], hazardous features [5] or other properties, for example temperature obtained from a thermal camera. Each element will have an assigned weighting, allowing for the context of the risk score to be considered. Applying more weight to elements that are more relevant in a given situation allows the proposed framework to be extendable as required. An overview of this framework is shown in Figure 3.

3.3. Hazardous Feature Descriptor - 3D Voxel HOG

The application of three dimensional descriptors to identify the properties of an object is a new concept, identifiers are required that allow us to differentiate hazardous objects from safe. A novel classification problem of recognising sharp and pointed areas in a scene (hazard features) is introduced. The



(a) (b) (c) (d)

Fig. 5: Example 3D Voxel HOG features from the tip of a knife blade and face of block, (a,e) visualised on its object in 3D, (b,f) the same 3D representation in two different orientations, (c,g) as a 2D Histogram and (d,h) as a 162 dimension feature vector

overall proposed classification approach for hazard areas in a scene is shown in (Figure 4). To achieve this 3D Voxel HOG is introduced which extends the original Histogram of Oriented Gradients for use in the third dimension. 3D VHOG is suitable for recognition of local shape characteristics and additionally considers an objects' density.

The traditional HOG uses the normalized combination of gradient vectors from a given number of pixels to build up a histogram of binned angles that relate to the feature. This is extended into the third dimension though the use of voxels and 2D histograms. Initially the voxel volume up is divided into set features spaces f comprised of a number of cubic 3D cells c, made up of voxels v. For each voxel within a cell the filter mask $[-1, 0, 1]$ is applied in all three dimensions giving us the gradient vector $\vec{g}$.

The magnitude $\|\vec{g}\|$ of the gradient vector is obtained and its orientation expressed using azimuth θ and zenith ϕ angles.

$$(\theta, \phi) = (\cos^{-1}(\frac{g_z}{\sqrt{g_x^2 + g_y^2 + g_z^2}}), \tan^{-1}(g_y, g_x)) \qquad (2)$$

A weight w is defined for each voxel, which is used to scale its contribution to its cell's 2D histogram. This is given by the mean value of the voxels within a given three dimensional kernel indicating the density over this area. By applying this weight, the proposed approach provides accurate estimates also in the presence of noise. Due to the local nature of the proposed feature, issues related to the normalization of a mesh are avoided, removing a potentially complex pre-processing step.

Once these values are established the voxels within each cell are binned into a 2D histogram h according to their θ and ϕ angles. The value added to a bin is given as the weighted magnitude of the vector $w\|\vec{g}\|$. Finally all cell histograms within a feature h_f are normalised using the L_2 norm and finally vectorised.

$$h_f \rightarrow \frac{h_f}{\sqrt{\|\vec{g}\|_2^2 + e^2}} \qquad (3)$$

$$\vec{x}^{3DVHOG} = \{h_{1,1}, ..., h_{1,\varphi}, ..., h_{\theta,\varphi}\} \qquad (4)$$

When visualised, these 2D histograms present a way to identify differing features within an object (Figure 5c). Another form of visualisation plots each possible gradient vector within local 3D histograms, showing the most common gradient vectors as stronger (Figure 5a).

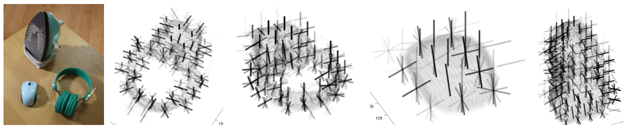


Fig. 6: Example 3D Voxel HOG visualisations of objects from the 3DRS Dataset.

These feature vectors are used to create a trained model that unknown shape features can be tested against. A binary classification is returned defining the object as either being hazardous or not. Adaboost is a learning technique that creates a non linear classifier to separate data into two groups. Weak classifiers are defined with a final strong classifier being a combination of these. At each iteration the weak classifiers with the lowest error margin are used to define the next in a 'greedy fashion'. Regarding the proposed features in both cases given N training examples $(\vec{x}_1, ..., \vec{x}_N)$, the corresponding labels $(y_1, ..., y_N)$ with $y_i \in \{-1, 1\}$, and an initial distribution of weights $W_1(i)$ a strong classification model $H(x)$ is obtained based on the weak classifiers h. The weak classifiers are trained over a number of iterations Q using the weights' distribution W_t. In each iteration the error ϵ_t is estimated based on the current weights W_t, that are updated before the next iteration.

$$W_{t+1}(i) = W_t(i)\exp(-a_t y_i h_t(x_i))/Z_t \qquad (5)$$

where $a_t = -\frac{1}{2}\log(\epsilon_t/(1-\epsilon_t))$ and $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$ is a normalization factor. The strong classifier is defined as $H(x) = sign(f(x))$, where $f(x) = \frac{\vec{a}\cdot\vec{h}(x)}{\|\vec{a}\|_1}$.

Regarding the boosting approach, because of the way weak classifiers are selected a complicated feature problem can be broken down and classified using a sparse classification rule, based on only a few features. This makes computation much faster as only a subset of the features is used. This is essential if the methodology is to be implemented in a real time scenario. Another advantage of this approach is the explicit minimisation of error, whilst implicitly maximising the margin. This ensures the final strong classifier is general avoiding the problems of overfitting.

Finally, in order to define the 'hazard features' element $E = D^{3DVHOG}$ of the risk score R in equation (1) the obtained outcomes from the classification process above are utilised.

$$D^{3DVHOG} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{k=1}^{M} w_D G(j,k)}{\sum_{k=1}^{M} G(j,k)}\right)$$

$$D^{\omega} = \frac{1}{m}\sum_{j=1}^{m} w_D(j) \qquad (6)$$

where $w_D = f(x)$ normalised and $G = \frac{1}{2}(sign(f(x)) + 1)$.

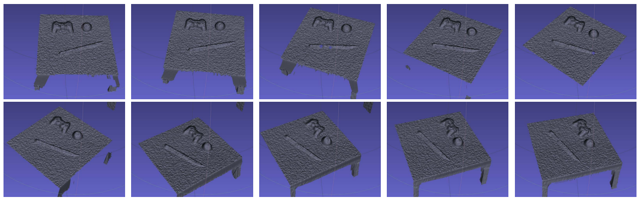


Fig. 7: Example frame sequence of a table top scene, starting top left and moving to the right over time.

4. RESULTS

4.1. Experiment Environment

To effectively test this problem we analyse scenes comprising of household objects placed on a table from the 3DRS Dataset. A scene is defined as a fixed environment from which a series of sequential frames are recorded. For the duration of the recording the recording device's angle and position relative to the scene is adjusted (Figure 7). 3 Scenes each with 10 frames were captured, aligned and voxelised to a resolution of 256 cubic voxels according to Section 2. The objects within the scene are categorised as 'hazardous' (bread knife, kitchen knife, iron, claw hammer) and 'safe' (ball, mug, game controller, condiment bottle). Each scene contains three objects; importantly objects have been chosen that have varying heights, this is to ensure that the artifacts created due to the depth shadows behind objects are included in the dataset. This ensures that the data used is challenging to the methodology. As each frame represents the scene from a different angle the depth shadow cast by the objects changes from frame to frame. Additionally no effort is made to derive context between frames, ensuring the system does not reply on any temporal knowledge of the scene to make an analysis.

4.2. Hazard Features Evaluation

Here we evaluate the risk level of each frame using an object's 'hazard features'. Using the 3D VHOG descriptor, features for each frame in each scene were defined and the groundtruth manually labeled at a cell level. For a comparison the features for another state-of-the-art 3D descriptor was also extracted and evaluated. In the proposed 3D Voxel HOG method a number of variables are defined. Based on experimental results the values for a feature block and cell size were set; 2 cubic cells and 16 cubic voxels respectively. The bins for the 2D histogram were set at 18 for θ and 9 for ϕ.

Once extracted, the histogram data from each feature (8 cells) was arranged into a mean 162 dimension feature vector for training using Adaboost. Training was done on one frame from each scene and a single model was used to evaluate the remaining 27 frames. The results for all the descriptors are summarised in table 1. Results are defined at an object level. One of the most important aspects of a risk evaluation system is that a hazardous object is not falsely classified as safe. Therefore it is essential that the sensitivity of the classifier is as high as possible.

	Feature	F1	Sensitivity	Accuracy
Scene	3D HOG	0.783	1.000	0.815
1	FAST IM	0.600	1.000	0.556
Scene	3D HOG	0.900	1.000	0.926
2	FAST IM	0.667	0.500	0.667
Scene	3D HOG	0.824	1.000	0.778
3	FAST IM	0.824	1.000	0.778

Table 1: Feature comparison against other existing 3D descriptors.

Object	B	C	K	Bk	M	I	H	Cb
3D VHOG	0.2	0.3	1	1	0.1	0.6	1	0.3
FASTIM	0.3	0.9	1	1	0.1	1	0.9	0.2

Table 2: Risk Score calculated for each object used in the test scenes

Figure 8 demonstrates that regardless of the features utilised the F1 scores stay consistent throughout the duration of the sequence indicating that accuracy is not lost over time.

Since the 'hazard' features of the testing objects have been estimated based on the proposed 3D VHOG descriptor and the classification mechanism, the level of risk based on the objects' characteristics is obtained using the equation 6. The obtained results using equation 1 for the tested objects are shown in table 2 indicating that the proposed method provides reasonable and accurate estimates.

5. CONCLUSIONS

In this work the concept of risk analysis is considered for 3D scenes in real time. The 3D Voxel HOG descriptor is utilised to represent the intrinsic properties of the objects and compared with other state of the art features, provided the highest accuracy. The ability to define hazardous objects through the use of the Risk Estimation framework has been demonstrated. The level of accuracy achieved indicates that there is great potential for the measurement of risk in real time applications.

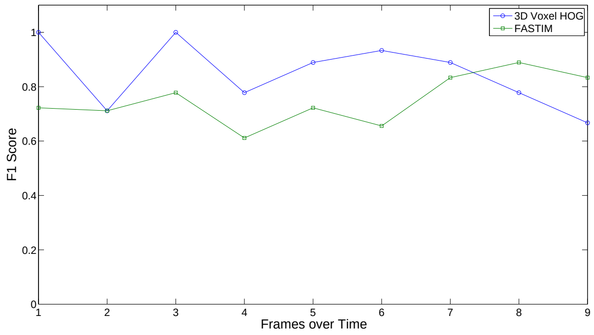


Fig. 8: Average F1 Score of each tested feature per frame across all scenes

6. REFERENCES

[1] N. Buch, S A Velastin, and J. Orwell, "A Review of Computer Vision Techniques for the Analysis of Urban Traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 920–939, Sept. 2011.

[2] Agnes Swadzba, Niklas Beuter, Sven Wachsmuth, and Franz Kummert, "Dynamic 3D scene analysis for acquiring articulated scene models," *2010 IEEE International Conference on Robotics and Automation*, pp. 134–141, May 2010.

[3] Liming Chen, Chris D Nugent, and Hui Wang, "A Knowledge-Driven Approach to Activity Recognition in Smart Homes," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 24, no. 6, pp. 961–974, 2012.

[4] Bo Zheng, Yibiao Zhao, Joey Yu, Katsushi Ikeuchi, and Song-Chun Zhu, "Scene Understanding by Reasoning Stability and Safety," *International Journal of Computer Vision*, vol. 112, no. 2, pp. 221–238, 2015.

[5] R Dupre, V Argyriou, G Tzimiropoulos, and D Greenhill, "3D Voxel HOG and Risk Estimation," in *IEEE DSP Singapore*, 2015.

[6] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, "Real-time human pose recognition in parts from single depth images," *Cvpr 2011*, pp. 1297–1304, June 2011.

[7] Y Freund and RE Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," *Computational learning theory*, 1995.

[8] Erik Stone and Marjorie Skubic, "Evaluation of an Inexpensive Depth Camera for Passive In-Home Fall Risk Assessment," *Proceedings of the 5th International ICST Conference on Pervasive Computing Technologies for Healthcare*, 2011.

[9] Ahmed Nabil Belbachir, Aneta Nowakowska, Stephan Schraml, Georg Wiesmann, and Robert Sablatnig, "Event-driven feature analysis in a 4D spatiotemporal representation for ambient assisted living," *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1570–1577, Nov. 2011.

[10] Bo Zheng, Yibiao Zhao, Joey C. Yu, Katsushi Ikeuchi, and Song-Chun Zhu, "Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3127–3134, June 2013.

[11] Zhaoyin Jia, Andrew Gallagher, Ashutosh Saxena, and Tsuhan Chen, "3D-Based Reasoning with Blocks, Support, and Stability," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2013.

[12] Andrea Fossati, Helmut Grabner, and Luc Van Gool, "Exploiting Physical Inconsistencies for 3D Scene Understanding," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*. IEEE, 2012, pp. 136–143.

[13] Sanjeev J Koppal and Srinivasa G Narasimhan, "Appearance derivatives for isonormal clustering of scenes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 8, pp. 1375–1385, 2009.

[14] D.C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2136–2143, June 2009.

[15] D Engel and C Curio, "Towards robust scene analysis: A versatile mid-level feature framework," *Poster presented at 4th International Conference on Cognitive Systems (CogSys)*, 2010.

[16] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. I, pp. 886–893.

[17] David Forsyth, "Object detection with discriminatively trained part-based models," *Computer*, vol. 47, no. 9, pp. 6–7, Sept. 2014.

[18] N. Buch, Mark Cracknell, J. Orwell, and S A Velastin, "Vehicle localisation and classification in urban CCTV streams," *Proc. 16th ITS WC*, pp. 1–8, 2009.

[19] Maximilian Scherer, Michael Walter, and Tobias Schreck, "Histograms of oriented gradients for 3d object retrieval," *Proceedings of the WSCG*, 2010.

[20] Shuai Tang, Xiaoyu Wang, Xutao Lv, Tony X. Han, James Keller, Zhihai He, Marjorie Skubic, and Shihong Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Computer VisionACCV*, 2013, vol. 7725 LNCS, pp. 525–538.

[21] Afzal Godil and AI Wagan, "Salient local 3D features for 3D shape retrieval," *IS&T/SPIE Electronic Imaging*, , no. Shrec, 2011.

[22] Ivan Sipiran and Benjamin Bustos, "Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes," *The Visual Computer*, vol. 27, no. 11, pp. 963–976, July 2011.

[23] Shahram Izadi, Andrew Davison, Andrew Fitzgibbon, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Dustin Freeman, "Kinect Fusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, 2011, p. 559.

[24] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz, "Aligning point cloud views using persistent feature histograms," *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pp. 3384–3391, 2008.

[25] A Trevor, Suat Gedikli, Radu Bogdan Rusu, and Henrik I. Christensen, "Efficient organized point cloud segmentation with connected components," in *Proceedings of Semantic Perception Mapping and Exploration*, 2013, pp. 1–6.

[26] Jian Huang Jian Huang, Roni Yagel Roni Yagel, Vassily Filippov Vassily Filippov, and Yair Kurzion Yair Kurzion, "An accurate method for voxelizing polygon meshes," *IEEE Symposium on Volume Visualization (Cat. No.989EX300)*, pp. 119–126,, 1998.

[27] Bo Zheng, Yibiao Zhao, and Jc Yu, "Detecting Potential Falling Objects by Inferring Human Action and Natural Disturbance," *IEEE Int. Conf. on Robotics*, 2014.