Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-15 (2022) 160–165

Next best view estimation for volumetric information gain

Alexandru Pop * **Levente Tamas** *

* *Technical University Cluj-Napoca, Automation Department, Romania*
e-mail: {Alexandru.Pop, Levente.Tamas}@aut.utcluj.ro

Abstract: In this work we propose a novel next best view (NBV) generation algorithm for volumetric information maximization. The primary data source is a Time-of-Flight (ToF) camera and the output is the next position of the depth sensor that maximizes a chosen score, either coverage or histogram of volumetric estimations for parallelepipedic shapes on the observed scene. Our learning-based method was validated on a large scale of real and synthetic data. The demo code, custom data sets, and videos are available on the author's website.

Keywords: Trajectory and path planning, perception and sensing, deep neural network

1. INTRODUCTION

The problem of reconstructing or mapping an a priori unknown space lies on the border of robotics and computer vision domains. This problem combines the perception of a scene with planning to gain the most information in consecutive measurements. To do this, the choice of sensor placements or viewpoints is essential, which in (Eidenberger and Scharinger, 2010) is mentioned as active perception whereas in (Mendoza et al., 2020) it is called next-best view (NBV) planning. The main problem addressed with NBV estimation is the finding of the optimal view sequence to minimize the number of measurement steps and to maximize in the same time the gained information. Thus, a well-tuned algorithm can determine in a few captures the whole 3D model of the unknown space or object, resulting in a quick reconstruction setup.

NBV estimation in robotics has been well studied in the last decades (C.I.Connolly, 1985), (Banta et al., 2000), (Massios et al., 1998), (Isler et al., 2016a), (Kriegel et al., 2015) and is still a research focus today (Vasquez-Gomez et al., 2021), (Zeng et al., 2020b), (Mendoza et al., 2020). The most common approach is based on the generation and testing of 2D or 3D sensors as presented in (Scott et al., 2003). In this approach, the NBV algorithm has as input a partial scene view, which in the case of a 3D sensor typically is a point cloud, some viewpoint candidates, and generates the next viewpoint for the sensor, to complete the previous measurements. For 3D sensors, the volumetric approach is the most popular (Zeng et al., 2020a), which transforms discrete 3D points into a voxel representation and, using advanced ray casting algorithms, predicts coverage and NBV. However, with the growth of the scene, this approach tends to be computationally demanding, so alternative solutions based also on deep learning solutions were proposed by Vasquez-Gomez et al. (2021), Zeng et al. (2020b), Mendoza et al. (2020).

Deep learning-based solutions are a data-driven approach to the NBV estimation problem, resulting in a fast heuristic solution even for 3D data, as proposed in (Mendoza et al., 2020). This paper specifically focuses on the problem of 3D volumetric estimation of a parallelepipedic object in a prior unknown space. More specifically, we are interested in the NBV, which reduces the uncertainty in the volume estimation of a regular geometric shape on the scene, as is visible in Figure 1.

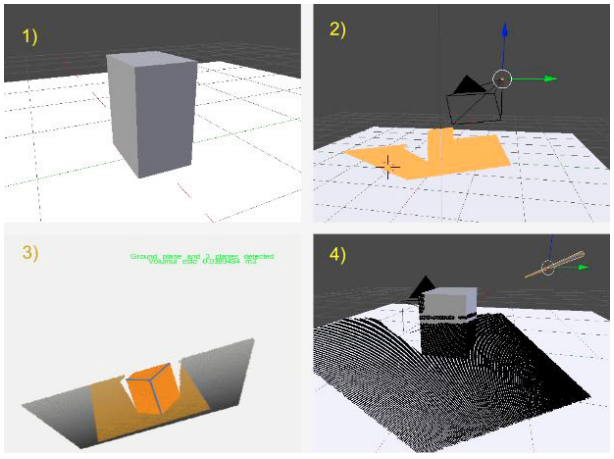

Fig. 1. Scan and NBV estimation for box-like objects

We assume that the only input to the network is a dense point cloud from a Time of Flight (ToF) type camera, without this being a hard constraint on the camera input type. We feed this input directly into the customized PC-NBV (Zeng et al., 2020b) network, in which we encapsulated the *volumetric information gain (VIG)*, i.e. the quality of the volumetric information about the regular geometric shape on the scene. Based on this information, we generate an NBV for the view which reduces the most the uncertainty in the volume estimation. This is relevant especially in a situation when from a

⋆ This work was financially supported by the Romanian National Authority for Scientific Research, CNCS-UEFISCDI, project number PN-III-P2-2.1-PTE-2019-0367 and PN-III-P3-3.6-H2020-2020-0060 and European Union's Horizon 2020 research and innovation programme under grant agreement No. 871295. Beside this, the NVidia and Analog Devices supported with hardware donation this work.

10.1016/j.ifacol.2022.07.625

single view the volume estimate is computed with low confidence. The choice of these box-like objects is well motivated by the clear need in the shipping and material handling applications for a volumetric estimate of boxes. Often these volumetric estimates are possible only from multi-view systems, thus the need for an efficient NBV is straightforward.

To generate the NBV constrained with VIG, we considered for the network training phase a large synthetic and real data set with ToF cameras that capture box-shaped objects, as can be seen in Figure 2. From our data set using advanced point cloud processing pipelines we extracted the box shapes based on custom planar surface ensembles as well as advanced corner detection and refinements similar to (Sommer et al., 2020), (Pop et al., 2021). To our knowledge, this is the first work to treat NBV in the context of VIG for box-like objects in unknown scenes.

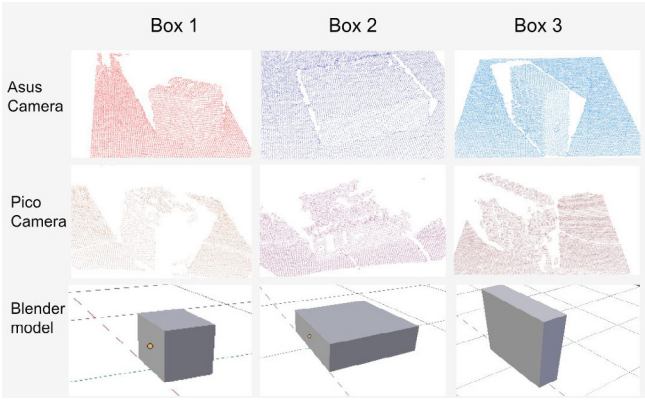

Fig. 2. Real camera outputs from Asus and Pico ToF cameras and synthetic box models used in Blensor

In addition to the standard NBV for generic point clouds, we considered the custom case of ToF, i.e. we incorporated in the VIG the ToF image specific constraints, such as the incidence angle of the camera light projection on the planar surfaces of the box which may result in measurement artifacts, or the closeness of the box to image boundaries which relates to the completeness of the observed box. In addition to this, we validated our algorithm on a large scale of real and synthetic data, focusing on the generalization capabilities of the proposed solution. The method proved to be able to generalize, i.e. for objects not used at the training phase was still able to predict correctly the NBV for volumetric estimation purposes.

The contribution of this paper is summarized as follows: 1) adaptation of the next best view problem to the volumetric information gain setup; 2) computation of volumetric information using planar and corner-based methods both on synthetic and real ToF data; 3) extension of the VIG setup for the training data set with ToF specific constraints such as the angle of the planar surfaces with the camera axes or the completeness of the measured object; 4) extension of the PC-NBV (Zeng et al., 2020b) with the generic case without view.

2. RELATED WORK

In this section, we shortly summarize the two main approaches for the NBV from the main literature: the traditional sampling-based ones and the learning-based variants using deep learning techniques.

2.1 Traditional methods

The traditional methods for finding the next best view can be classified into two categories: synthesis methods and generate-and-test methods (Zeng et al., 2020b). Synthesis methods compute directly the next best view under system constraints and although they are quick, they do not have the necessary robustness. The generate-and-test methods optimize a pipeline in which, for a given point cloud as input, the algorithm determines a set of viable candidate positions and, for each position, it determines a score. The selected position with the highest score is the NBV. The difficulty of the problem stems from the fact that the points in the different candidate positions are not known. As such, the algorithm uses the input point cloud to estimate a volumetric gain for each position. The final purpose of the NBV in the traditional methods falls in one of the following four classes: object or scene reconstruction (Tamas and Goron, 2014), object recognition (Tamas and Jensen, 2014; Tamas and Cozma, 2021) and pose estimation (Frohlich et al., 2021), (Blaga et al., 2021).The general outline of the pipeline for the traditional generate-and-test methods can be summarized in Figure 3.

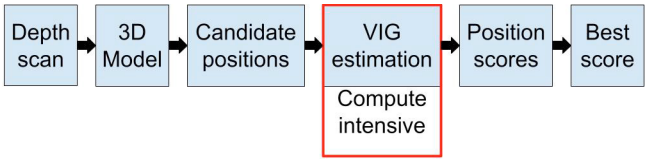

Fig. 3. Structure of the generate-and-test NBV pipeline

According to the objective for the Next Best View and the representation of the 3D model, the information gain algorithm can be constructed and each view can be scored. (Scott et al., 2003) introduced an important distinction between model-based vs. non-model-based reconstruction methods. Model-based methods assume a priori information about the object whereas non-model based do not use any information about the object geometry.

For object reconstruction applications, the discussion was started by C.I.Connolly (1985). His proposed method relied on a voxel space representation and used candidate positions from a spherical distribution around the object. The voxel space is necessary to determine the information gain of each view. Massios et al. (1998) introduced a quality criterion for views, while Banta et al. (2000) developed three criteria to select views based on occlusions. Vásquez-Gómez et al. (2009) introduced a utility function for the views and developed in (Vásquez and Sucar, 2011) a method for dealing with position error. In (Vasquez-Gomez et al., 2014) the two previous methods were combined to create a system capable of predicting views for reconstruction with position errors. Isler et al. (2016b) synthesized several volumetric gain algorithms in the voxel space. Delmerico et al. (2018) built on the findings of

Isler et al. (2016b) to create a comparison between the volumetric gain algorithms. A recent survey on traditional methods can be found in (Zeng et al., 2020a).

2.2 Learning based methods in 3D

Learning-based methods aim to avoid lengthy computations for each candidate view and instead try to estimate the scores for each view directly from the input 3D model representation. The neural network creates a signature for each depth image and manages to associate an array of scores that represent the estimates of volumetric gain for each view. If the network is successful, the scores for each view will be accurate enough to create a clear ordering of the views and the view with the highest score is the NBV, as is shown in Fig. 4.

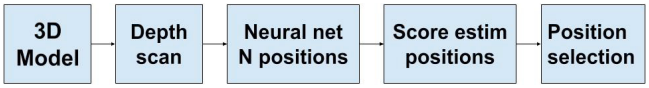

Fig. 4. Neural network architecture for NBV pipeline

Wu et al. (2015) brought an important contribution in this field with the neural network. The network can recognize and complete point cloud models, and, under this interpretation, the view that best completes the point cloud is the one selected. Another important contribution comes from Hepp et al. (2018). The networks learn to score views by utility, which makes it possible to choose a new position with the maximum estimated score. Their work is focused on scene exploration, rather than reconstruction. (Mendoza et al., 2020) is closest to our problem definition. In this work, the authors attempted to solve the NBV problem for reconstruction by converting point clouds to voxels and using a convolutional neural network to predict the view that maximizes the potential number of new voxels that can be added to the input point cloud. This method predicts the final position from a limited set of discrete positions, hence the problem is a classification one.

3. PROPOSED METHOD

Our approach was based on a deep learning active perception setup called PC-NBV by Zeng et al. (2020b) which is built on a multi-stage neural network. In essence, our network architecture is the same as PC-NBV but we do not use the view state vector, meaning we do not know the position of the camera. Furthermore, our network is built to predict one step without taking into consideration previous steps, as PC-NBV does with the view state. Without the position of the camera, our network can be immediately deployed without needing pose from IMU information.

We consider a set of N possible camera placements with a corresponding N-dimensional array of scores, one score for each possible view. In the PC-NBV architecture, the possible positions of the camera are considered in order, hence the first element of the score array corresponds to the first possible position, the second score corresponds to the second position, and so forth. For our network, we do not know a priori which of the possible positions the camera is in. We only know that there is an overlying order between them. Furthermore, we consider point clouds from a camera position to have clear differences from point clouds of different positions. Considering the score of the current camera position as the first position in the output array, we can consider in order the rest of the camera positions that follow the current one. By applying a circular permutation to the output scores where we know the global position, we create training data for the neural network to predict the permutation according to the input point cloud, as can be seen in Fig. 5 and Fig. 6.

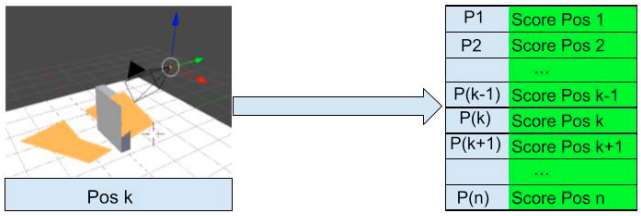

Fig. 5. Network predictions without permuted scores

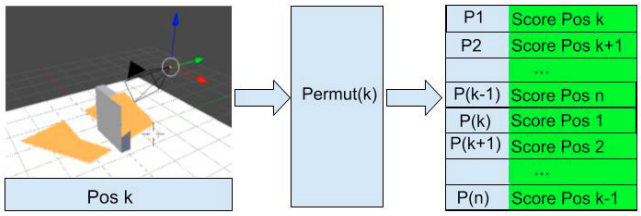

Fig. 6. Network predictions without permuted scores

The PC-NBV network and our modified network can be trained to predict scores for each view. This means that we select scores that reflect a continuous function based on the input point cloud, or we select scores that reflect a discrete choice between the views, like one hot encoder. Our modified network has to compensate for the lack of camera position. The scores for which the PC-NBV network was built are related to coverage, i.e. the completeness of the 3D reconstruction obtained for each position in a limited set. By changing the score to reflect a different aspect of the point cloud, the network can be trained to choose the views that maximize a different objective.

To test the different training methods, we used three networks with different training: **1) PC-NBV-noview-coverage**, **2) PC-NBV-noview-vol-real**, and **3) PC-NBV-noview-vol-synthetic**. PC-NBV-noview-coverage is a direct comparison between PC-NBV and our modified network without the view state regarding point cloud coverage. PC-NBV-noview-vol-real outputs a one-hot encoding of the position with the lowest volumetric estimation error on real data. PC-NBV-noview-vol-synthetic is the same as PC-NBV-noview-vol-real but it is trained on synthetic data. All our networks use the same architecture detailed in Fig. 7 but were trained with different data.

3.1 Architecture details

The network is composed of parts from different neural network architectures that were designed for tasks related to extracting features from point clouds. The backbone of the network is provided by the feature extraction unit described by Yuan et al. (2018). This permits the computation of features from the point cloud and the selection

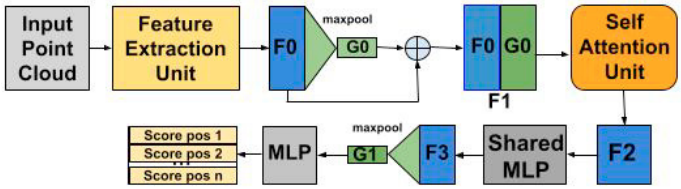

Fig. 7. General structure of the PC-NBV network without view state vector

of a global feature using max pooling. This network architecture was further enhanced with the feature extraction module described by Yifan et al. (2019), the self-attention module described by Zhang et al. (2019) and utilized by Li et al. (2019), with the final multilayer perceptron used to link a final score to the features. All these network modules are synthesized in the PC-NBV network described by Zeng et al. (2020b).The self-attention module is described in (Zhang et al., 2019) and a version of it is implemented in the network of Li et al. (2019) and subsequently by Zeng et al. (2020b). The goal of this module is to help compute long-distance interactions between features in the point cloud and to enhance feature integration after concatenation.

3.2 Loss function

The loss function used in all three networks is the Mean Squared Error (MSE) between the ground truth scores and the predicted scores, same as in (Zeng et al., 2020b). For the first network, the MSE is calculated between the circularly permuted coverage scores and the network output. For the second and third networks, the MSE is calculated between the permuted position selection array and the network predictions.

The network is trained with the same parameters as suggested by Zeng et al. (2020b). The starting learning rate is 0.0001 with a batch size of 32. The learning rate decays every 50000 iterations by 0.7. $\lambda = 0.0001$ is chosen for the loss L_2 utilized in weight regularization.

3.3 Volumetric information gain (VIG)

For volumetric information gain (VIG), we used to rank potential views with two different approaches. Each approach has a different objective for the views; hence, they are used for different purposes. Coverage based VIG is used for the reconstruction problem,where the goal is to gather in each successive view the most number of points to reconstruct an object whereas the VIG based on volumetric estimation error is an example of an alternative score for the views that reflects a different objective, namely the improvement of the volumetric measurement.

Coverage based VIG The coverage represents the percentage of points from the ground truth which have at least one neighbouring point at a distance bellow a threshold. For a sufficiently low threshold, each new view offers new points that will be close to the ground truth, eventually leading to coverage close to 95-99%. Because each successive new view offers less and less new points, the first views are the most important. To simplify the problem, we only considered the first move; therefore, the network must choose between multiple possible views, the one that offers the highest coverage score.

VIG based on volumetric estimation error Histogram The second method involves finding the position for which there is a higher probability of having a volumetric estimation close to 0. The volumetric estimation is done using the algorithm described by Sommer et al. (2020). Sommer's algorithm outputs a collection of lines obtained from projecting perpendicular planes. By selecting line triplets that are close to each other, the width, length, and height of a box can be identified and the volume is computed. Knowing the ground truth volume values, the relative volumetric error can be determined for each volume estimate. The relative volumetric error represents the value used to compute the NBV.

A comparison between volumetric estimation error histograms at each position is needed to compute the score by which the views are chosen. The view with the best histogram score will be chosen. Thus for each object type we must sample a large enough number of point clouds in each view and to obtain the volumetric estimation error for each of the sampled point cloud. By scoring the histograms of each position, we can select the position with the highest histogram score as the NBV. Thus the VIG is given by the score associated to the histogram of potential volumetric errors for each position, as can be seen in Fig. 8.

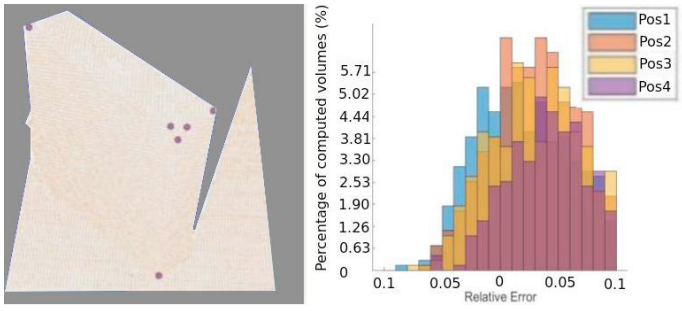

Fig. 8. Volume Information gain estimation and histogram

Three datasets were created in order to train and test the PC-NBV network. **PC-NBV-noview-coverage** was used to show that the PC-NBV network could learn to predict the scores permuted circularly by the starting position, as can be seen in Fig. 10. This modification would exclude the need for the view state from the network architecture and would allow for the direct use of a point cloud as input. For each point cloud, the output scores must be computed and associated. For the first network, the output scores are calculated with the coverage finding algorithm described by Zeng et al. (2020b). The scores are then permuted circularly by the starting position.

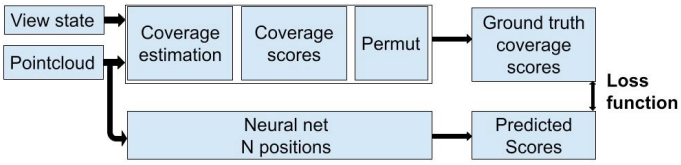

Fig. 9. Training for PC-NBV-noview-coverage

PC-NBV-noview-vol-real is trained with point clouds from a real camera whereas **PC-NBV-noview-vol-synt**

is trained with point clouds from synthetic data. Each network is optimized to predict a selection score that represents the position that minimizes the volumetric error. The working structure can be seen in Fig. 9. Since the selection score is built by analyzing the scores from each view, a histogram with the volumetric estimations from each position is needed to create the training data. The volume of the box is estimated using the orthogonal planes for each training point cloud. Knowing the ground truth dimensions of the box, the relative estimation error is computed for each box. For each position of the box, a histogram is made with the error estimations for all point clouds in the set. Once all histograms have been determined for the training and validation set, the position with the largest number of observations is selected.

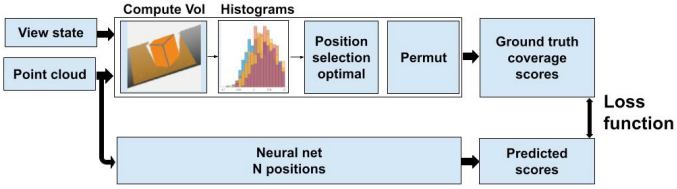

Fig. 10. PC-NBV-vol-synth and PC-NBV-vol-real modules

4. EVALUATION

For predicting permuted scores, the network was compared with PC-NBV on the same test data set. The network needs to be able to predict permuted coverage scores without having the view state as input. If the scores are similar to PC-NBV, the proof of concept of the network without the view state is complete.

For predicting permuted position on real data, the network is evaluated using point clouds from the same recording. The network needs to match the point clouds with the right permuted position for each box.

Finally,two data sets are used for synthetic point clouds with unknown camera position. The first set consists of point clouds taken for the same boxes from the same positions with slight random positions and rotation errors for the camera. The second set consists of point clouds from different boxes in the same positions with small random errors.

4.1 Dataset used for evaluation

PC-NBV-noview-coverage The training instructions mentioned by Zeng et al. (2020b) were used with the Shapenet database (Chang et al., 2015) being selected. 4000 models from 8 classes of objects were selected for training, 800 different models from the same eight classes were used for validation, and 400 objects from eight other classes were used for testing.

PC-NBV-noview-vol-real 3 boxes as shown in figure 2 were captured using 2 ToF cameras from 4 different positions and angles. The resulting point clouds were randomly split into 70% train, 30% valid.

Synthetic data PC-NBV-vol-synth 11 box models with different dimensions were created in Blensor, with a set of four predetermined positions was used together with random positioning and rotation to the camera to induce differences between the resulting point clouds in the same position. 700 scans/position/box were taken for training and 300 scans/position/box for validation.

4.2 Performance evaluation and comparison

Predicting coverage scores without view state As can be seen in Table 1, the evaluations of the first network were compared with the one from PC-NBV with view state. Our network which is designated NBV-noview-coverage, uses only the point cloud as input and outputs the scores of neighbouring views. The percentage of network predictions that matched ground truth data is given in the exact position column (exact.pos (%)). Since different views can give similar scores, we considered the percentage of network predictions given in column Approximate positions (Ap. pos)(%) that have a coverage score less than the greedy score by a maximum threshold given in column 2 of Table 1 as Cov. diff. The missing camera position leads to a slight worsening of the performance but leads to greater use-case flexibility. For a 6 % drop in accuracy, the network can only use the point cloud as input and can be used directly with a camera without needing IMU data to determine the position and orientation of the camera.

Table 1. Coverage-Based VIG

Model	Cov. diff	Fw. time (s)	Ap. pos. (%)	Exact pos. (%)
NBV-view-cov.	0.05	0.024	91.7	61.69
	0.03	0.024	84.28	61.69
	0.01	0.024	73.67	61.69
	0.005	0.024	69.19	61.69
NBV-noview-cov.	0.05	0.028	85.7	56.22
	0.03	0.028	77.94	56.22
	0.01	0.024	66.63	56.22
	0.005	0.024	62.32	56.22

Table 2. VIG based on histograms

Model	Time (s)	Acc. same box (%)	Acc. diff. box(%)
nv-vol-syn.	0.022	99.35	22.34
nv-vol-real	0.037	100	-

Predicting NBV using histograms VIG In Table 2, two networks that used VIG histograms are compared. Both networks have high accuracy for point clouds taken under the same conditions as the training data. In the case of a synthetic data network, the test data from the same box models have a high accuracy, but for the point clouds from different box models, the network output is less robust.

5. CONCLUSIONS

In this work, we presented a learning-based next best view generation algorithm for the volumetric estimation of regular geometric shapes from 3D point clouds. We first showed that the network is capable of learning permuted scores, allowing us to use only point clouds. Afterwards the networks trained with real and synthetic point cloud data can accurately predict the NBV for point clouds of the same object class as the training data. In the future, we intend to improve the prediction for point clouds taken from different object classes and to generalize our algorithm for arbitrary shapes in a scene.

REFERENCES

Banta, J.E., Wong, L.M., Dumont, C., and Abidi, M.A. (2000). A next-best-view system for autonomous 3-D object reconstruction. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, 30(5), 589–598.

Blaga, A., Militaru, C., Mezei, A.D., and Tamas, L. (2021). Augmented reality integration into MES for connected workers. *Robotics and Computer-Integrated Manufacturing*, 68, 102057.

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. (2015). ShapeNet: An Information-Rich 3D Model Repository.

C.I.Connolly (1985). The Determination of Next Best Views. 432–435.

Delmerico, J., Sabzevari, R., and Scaramuzza, D. (2018). A comparison of volumetric information gain metrics for active 3D object reconstruction. *Autonomous Robots*, 42(2), 197–208.

Eidenberger, R. and Scharinger, J. (2010). Active perception and scene modeling by planning with probabilistic 6D object poses. *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, 1036–1043.

Frohlich, R., Tamas, L., and Kato, Z. (2021). Absolute pose estimation of central cameras using planar regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2), 377–391.

Hepp, B., Dey, D., Sinha, S.N., Kapoor, A., Joshi, N., and Hilliges, O. (2018). Learn-to-score: Efficient 3D scene exploration by predicting view utility. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11219 LNCS, 455–472.

Isler, S., Sabzevari, R., Delmerico, J., and Scaramuzza, D. (2016a). An information gain formulation for active volumetric 3D reconstruction. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June, 3477–3484.

Isler, S., Sabzevari, R., Delmerico, J., and Scaramuzza, D. (2016b). An information gain formulation for active volumetric 3d reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 3477–3484. IEEE.

Kriegel, S., Rink, C., Bodenmüller, T., and Suppa, M. (2015). Efficient next-best-scan planning for autonomous 3D surface reconstruction of unknown objects. *Journal of Real-Time Image Processing*, 10(4), 611–631.

Li, R., Li, X., Fu, C.W., Cohen-Or, D., and Heng, P.A. (2019). PU-GAN: A point cloud upsampling adversarial network. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob(c), 7202–7211.

Massios, N.A., Fisher, R.B., et al. (1998). *A best next view selection algorithm incorporating a quality criterion*, volume 2. Department of Artificial Intelligence, University of Edinburgh.

Mendoza, M., Vasquez-Gomez, J.I., Taud, H., Sucar, L.E., and Reta, C. (2020). Supervised learning of the next-best-view for 3d object reconstruction. *Pattern Recognition Letters*, 133, 224–231.

Pop, M.L., Molnar, S., Pop, A., Kelenyi, B., Tamas, L., and Cozma, A. (2021). Cnn based tof image processing.

Scott, W.R., Roth, G., and Rivest, J.F. (2003). View planning for automated three-dimensional object reconstruction and inspection. *ACM Computing Surveys (CSUR)*, 35(1), 64–96.

Sommer, C., Sun, Y., Guibas, L., Cremers, D., and Birdal, T. (2020). From planes to corners: Multi-purpose primitive detection in unorganized 3D point clouds. *IEEE Robotics and Automation Letters*, 5(2), 1764–1771.

Tamas, L. and Cozma, A. (2021). Embedded real-time people detection and tracking with time-of-flight camera. In *Proc. of SPIE Vol*, volume 11736.

Tamas, L. and Goron, L.C. (2014). 3d semantic interpretation for robot perception inside office environments. *Engineering Applications of Artificial Intelligence*, 32, 76–87.

Tamas, L. and Jensen, B. (2014). Robustness analysis of 3d feature descriptors for object recognition using a time-of-flight camera. In *22nd Mediterranean Conference on Control and Automation*, 1020–1025. IEEE.

Vásquez, J.I. and Sucar, L.E. (2011). Next-best-view planning for 3d object reconstruction under positioning error. In *Mexican International Conference on Artificial Intelligence*, 429–442. Springer.

Vasquez-Gomez, J.I., Sucar, L.E., Murrieta-Cid, R., and Lopez-Damian, E. (2014). Volumetric next-best-view planning for 3D object reconstruction with positioning error. *International Journal of Advanced Robotic Systems*, 11.

Vasquez-Gomez, J.I., Troncoso, D., Becerra, I., Sucar, E., and Murrieta-Cid, R. (2021). Next-best-view regression using a 3d convolutional neural network. *arXiv preprint arXiv:2101.09397*.

Vásquez-Gómez, J.I., Löpez-Damian, E., and Sucar, L.E. (2009). View planning for 3d object reconstruction. In *2009 International Conference on Intelligent Robots and Systems*, 4015–4020. IEEE.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shapes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015, 1912–1920.

Yifan, W., Wu, S., Huang, H., Cohen-Or, D., and Sorkine-Hornung, O. (2019). Patch-based progressive 3D point set upsampling. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, 5951–5960.

Yuan, W., Khot, T., Held, D., Mertz, C., and Hebert, M. (2018). PCN: Point completion network. *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*.

Zeng, R., Wen, Y., Zhao, W., and Liu, Y.J. (2020a). View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 6(3), 225–245.

Zeng, R., Zhao, W., and Liu, Y.j. (2020b). PC-NBV : A Point Cloud Based Deep Network for Efficient Next Best View Planning PC-NBV. *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *Int. Conf. on machine learning*, 7354–7363. PMLR.