Efficient Recognition of Multiple 3D Objects in Point Clouds with a Multifoveation Approach

Fabio F. Oliveira, Luiz M. G. Goncalves, Marcelo A. C. Fernandes, Rafael B. Gomes
Department of Computer Engineering and Automation
Federal University of Rio Grande do Norte (UFRN)
Natal - RN 59.078–970 - Brazil
Email: fabio.veritate@gmail.com, {lmarcos, mfernandes}@dca.ufrn.br, rafaelbg@dimap.ufrn.br

***Abstract*—Technological innovations in the hardware of RGB-D sensors have allowed the acquisition of 3D point clouds in real time. Consequently, various applications have arisen related to the 3D world, which are receiving increasing attention from researchers. One of the main problems that remains is the demand for computationally intensive processing that is required by optimized approaches to deal with 3D vision modeling, especially when it is necessary to perform tasks in real time. A previously proposed multi-resolution 3D model known as foveated point clouds can be a possible solution to this problem. However, this is a model that is limited to a single foveated structure with context dependent mobility. So to recognize several objects one would need to move the fovea in the 3D data, sequentially. In this work, we propose a new solution this problem, providing data reduction for feature detection, by using a multifoveation in the point cloud. Notice that the simple application of several foveated structures results in a considerable increase of processing since there are intersections between regions of distinct structures, which are processed multiple times. Towards solving this problem, the current proposal brings an algorithm that avoids the processing of redundant regions, which results in even more reduced processing time. Such approach can be used to identify objects in 3D point clouds, one of the key tasks for real-time applications as robotics vision, with efficient synchronization allowing the validation of the model and the verification of its applicability in the context of computer vision. Experimental results demonstrate a performance gain of at least 27.21% in processing time while retaining the main features of the original, and maintaining the recognition quality rate in comparison with state-of-the-art 3D object recognition methods.**

I. Introduction

The main purpose and the proposal of this paper[1] is to extend the model based on the foveated point cloud (FPC) [1] coming up with a sampling approach to 3D data that reduces processing time while achieving multiple foveated point clouds (MFPC), at several points of interest. This extension allows using this approach to deal with, for example, the search of multiple objects in a scene instead of a single one as in the previous work, with data compression for server storage, and for assisting simultaneous localization and mapping (SLAM) tasks. Our efforts were started trying to accelerate the Visual SLAM [2] with the results achieved, we prepared the way for the development of the MFPC.

The contribution of this work is the approach to speed-up the processing on multifoveated point clouds by avoiding redundant computations in overlapping regions. The novelty in relation to previous work is that, instead of only one moving fovea, several ones can be used at a time without redundancy on the computations. The results of the experiments provided in this work show that this reduction brings benefits in the performance of some tasks being executed, such as recognition, speeding up the applications without affecting the precision. This makes possible the use of multifoveation to deal with the search of multiple objects in a scene, instead of just one object as in previous work [1]. Consequently, multiple targets in a scene can be addressed in robotics applications. Other visioned applications as 3D data transmission through the Internet where data reduction is mandatory can also use such approach.

Thus, the scope of the present model is restricted to visual data features capture with depth data using several types of structures in parallel, which reduces the amount of information without creating redundant information to streamline a task in dynamic or static environments. Thus, in the later sections we will have a brief discussion about the mechanism and its respective results.

[1]This work relates to a M.Sc. thesis.

II. Background Theory

To achieve high level of autonomy, it is necessary for a robot to react without human intervention to stimuli provided by the environment [3, 4, 5]. Robotic vision applications use several types of sensors to capture stimuli data. To better comprehend the approach to multifoveation in point cloud proposed in this paper that can be used for robotic vision, in this section, we present the basic ideas of multi-resolution and the previous foveation approaches, starting with the single foveation that has allowed performing real-time tasks mainly for robotic vision.

A. Multiresolution and Foveation

Our brain can acquire and abstract visual information quickly and efficiently deciding where to focus and what is the sequence of fixations [6]. This sequence of fixations is related to cognitive mechanisms controlled by our visual attention mechanism. It is understood that the biological system of human vision has two types of visual attention behavior called top-down and bottom-up [7, 8]. Top-down attention is an approach that refers to the internal orientation of care based on prior knowledge, desired goals, and planning. In contrast, bottom-up attention is directed purely by stimuli of external factors that stand out because of their inherent properties in relation to the background.

Thus, no matter which of the above tasks is performed, they need high processing rates, which makes vision one of the most complex activities for a robotic system implementation, since a single view requires a huge amount of data. Thus, using some technique for diminishing the amount of input data has proven to be a good idea [8, 9]. Many types of methods optimize image processing tasks. One of these is the representation based on the non-uniform density called foveation, which has as one of its aspects a biologically inspired model that mimics the mapping of the retina to the visual cortex to deal with this amount of data [10, 11].

The foveation process can be performed by means of software downsampling [10], with reduced sampling [12] for systems that use at least two cameras [13]. Each type of execution of the foveation process has its relevant advantages and disadvantages.

III. Multifoveated Point Cloud-Proposal

Our proposal in this paper is a natural evolution of the multi-resolution with centered fovea (MCF) [8], going through the moving fovea (MMF) [9] and is based in the FPC [1] with the ideas of models of multifoveation 2D [14, 15]. By grouping these two bases, it is proposed to create a model that reduces the processing time in the detection of objects, avoiding the processing of redundant regions generated by the overlap of the foveated structures, allowing a better performance in a multiple objects recognition task.

The idea is to apply several foveated structures on the PC, but with an additional step that remove the redundant points. In this scheme, it is possible to reduce the processing time while the densities around the several foveae are enough to guarantee the detection of the objects. In this scheme, the FS is a sectioned frame having different resolutions (multi-resolution), where the size of the level is inversely proportional to the quality of the resolution, so that each of these levels is successively encompassed by the other in such a way that all are disjoint. For further details all theoretical development in detail are available in our recently published work [16].

A. Approach to Cloud Multifoveation

The FPC proposal has as the main mechanism the down-sampling in the original point cloud using concentric boxes, where each one represents a level, producing a PC with a different density for each box. In the context of multiple foveated structures (MFPC), applying multiple moving foveae would result in a considerable increase in processing, since intersections between regions of distinct structures would be processed several times. In this way, our approach to the multifoveated model has as a specific case the proposal in the previous FPC. The foveated point cloud is achieved by downsampling the original PC using concentric boxes, each representing a level, which is the case for only one foveated structure. The operation of multiple structures without treatment is similar, although redundancy is eventually generated in regions to be processed, in addition to having more points than in the original scene.

IV. Proposed Object Recognition Scheme

In this section, we discuss the application of the proposed approach coming up with a new correspondence grouping algorithm, which is somewhat similar as the one presented by Tombari and Stefano [17]. The search is done only in the object-to-scene direction and not the scene to the object, allowing the system to find multiple instances of the same object in a single scene shown in Figure 1.

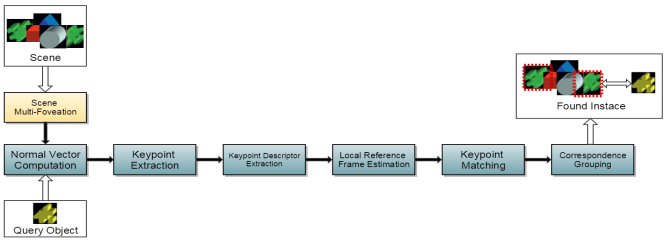

Fig. 1: The 3D object recognition algorithm based on matching: the proposal with the object recognition scheme 3D multi-foveation. The scene is downsampled through several foveated structures, considerably reducing the number of points to be processed without compromising overall accuracy, as described in the text.

A. Object Recognition in Multifoveated Point Clouds

The MFPC is applied to according to Figure 1 and the parameters described and defined in papers [16]. The estimation of the normal of each point can be made previous or after the foveation process. When choosing to make the estimation previously, computation is costlier, but the captured geometric traits of the scene are less distorted. However, we continue to choose to preserve the geometry of the scene, prioritizing the best accuracy.

To respect the multi-resolution of the scene PC, we use key-point extraction adapted to depend on the different and specific resolution levels in each FS, possibly differing from the sampling radius used $d_0, \ldots, d_m$. We also use the modified correspondence grouping algorithm that accommodate the extraction of the key-points in each point cloud of the multifoveation. The points of the scene are downsampling using several radii r_k for each k level and $k \in [0, \ldots, m]$, where all the foveated structures of the scene have the same numbers of levels. From the arbitrary determination of the extreme radii of the structures are the first level (level 0) having a radii r_0 and last (level m) using a distance r_m. The intermediate levels use linear interpolation for radii estimation, like what is done in the size of the boxes of the foveated structures.

Therefore, by adopting these radius downsampling approaches made in [1], considerable time savings can be achieved by reducing the number of key-points both in the computing of descriptors and in the matching step, resulting in the large increase in the density of the key-points near the fovea position without significant increase in the total number of original points in the scene. Thus, it improves the efficiency of the detection of objects if the foveated structures are defined correctly and reduces the number of false matches of descriptors.

The FPC model can only recognize several objects if all are close, considering a fovea box which does not cover much of the PC. In this new MFPC model, the objects can be identified without this restriction, since the distribution of multiple structures in the scene allow the proper positioning of each zone of maximum resolution.

V. Results

After deciding the types of strategies for comparison, we execute each of them. Two groups of graphs are used to display the information acquisitions. They are made up of four types that join the six approaches providing a better comparison between the strategies. The first group of graphs brings information that is related to the performance of the approaches: quantity of points, number of key-points, average time and number of objects found. The second group of graphs tries to explain the accuracy of the strategies in relation to the characteristics of the execution of the algorithm of object detection of Tombari and Stefano [17]. The characteristics of this group are: correspondence between points, amount of true-positive, precision and recall. For each scene, the two groups of graphs were used, besides elaborating tables which allow having a better precision in the execution times sampled of each scene for each strategy.

N°	Description
1	Non-foveated [17]
2	Multifoveated *raw*
3	Multifoveated with redundancy treatment (MFPC)
4	Foveated covering all object simultaneously (FCAS)
5	Foveated covering the rightmost object (FCRM)
6	Foveated covering the leftmost object (FCLM)

TABLE I: Enumeration of the treated approaches in the comparison of results for both the scenes the comparison experiments.

The related groups are shown in Figures 2 and 3, where we have the performance group and the accuracy group, respectively. The performance of our proposal was one of the concerns, since the recognition algorithm requires many computational resources in relation to 2D proposals. In the set of graphs referring to Figure 2, the ordinates axis is related to the strategies used mapped in Table I.

A. Results Analyzed from Performance Group

By analyzing Figure 2a, we can see that the number of points (abscissa axis) of the non-foveated (Experiment 1) is much more accentuated than the models with single fovea and multifoveated. The model non-foveated has the same total of points of the original scene that is 281,097 points, since the models that use the foveation are with quantities below 150,000 points. In our proposal (Experiment 3 (MFPC), we notice that it is the third largest decrease of points, losing only to the experiments that are involving only one object, which are the FCRM and FCLM. The multi-foveated raw model (Experiment 2) has a reduction of points generating redundancy of points, a fact that is explained by the arrangement of structures where only the lower density levels have intersections. Then, it is shown in Figure 2a that our strategy would have advantage over the others in the amount of points used to find all objects, using fewer points for the scene with more distributed objects.

N°	I_{lower}	μ	I_{upper}	σ^2	Exp. 3 (%)
1	0.765	0.776	0.787	4.0×10^{-4}	45.73
2	0.582	0.593	0.604	4.0×10^{-4}	28.96
3	0.415	0.421	0.427	1.3×10^{-4}	0
4	0.564	0.579	0.593	6.5×10^{-4}	27.21
5	0.302	0.305	0.308	0.3×10^{-4}	-
6	0.294	0.297	0.300	0.3×10^{-4}	-
6 + 5	-	0.596	-	-	29.98

TABLE II: Comparison of the execution times in seconds of the strategies performed and the reduction time in relation to experiment 3 in percentage. The confidence interval used is 95% by t-Student.

Looking at the results of key-points extraction based on each strategy (Figure 2b), it depends on the positioning of the fovea boxes for the foveated models. Notice that Experiment 4 has a much higher number of points compared to Experiment 3 that has a similar amount of extracted key-points. It is possible to preserve the descriptiveness of the scene with a significant reduction in the number of points. In the multifoveated raw, more key-points are found than in the non-foveated model. These key-points are duplicated or distorted in the scene generated by the redundancies, what affects the local descriptiveness by modifying it, being this the main problem of the approach multifoveated raw.

The expectation of the objects found is represented in the graph shown in Figure 2c. We can see that all strategies have expected results, except the strategy that uses the multifoveated raw. This fact can be explained initially using Figure 2b, since we have a higher number of key-points in relation to the non-foveated strategy with the same configurations. The scene distortion caused by the multifoveated raw has brought this false detection.

The performances of the execution times of the strategies used are shown in Figure 2d. The graph shown is constructed with 15 consecutive samples of execution times of each strategy, which allows to elaborate a confidence interval that guarantee a good estimation of the average, as seen in the Table II. As previously mentioned, Experiment 6 and 5 represent the strategy of using only one foveated structure to identify several objects at a time. The evaluations of the average times of performance of these two experiments are combined for comparison with the shortest average time (experiment MFPC) obtained in Table II reduction column. Table II has the following information: I_{lower} (lower end), I_{upper} (upper end), μ (arithmetic mean), σ^2 (variance) and reduction of time in relation to the experiment MFPC. The ends are maximum and minimum fluctuation of the acquired times of each experiment done.

We can see in Figure 2d that the best performance in

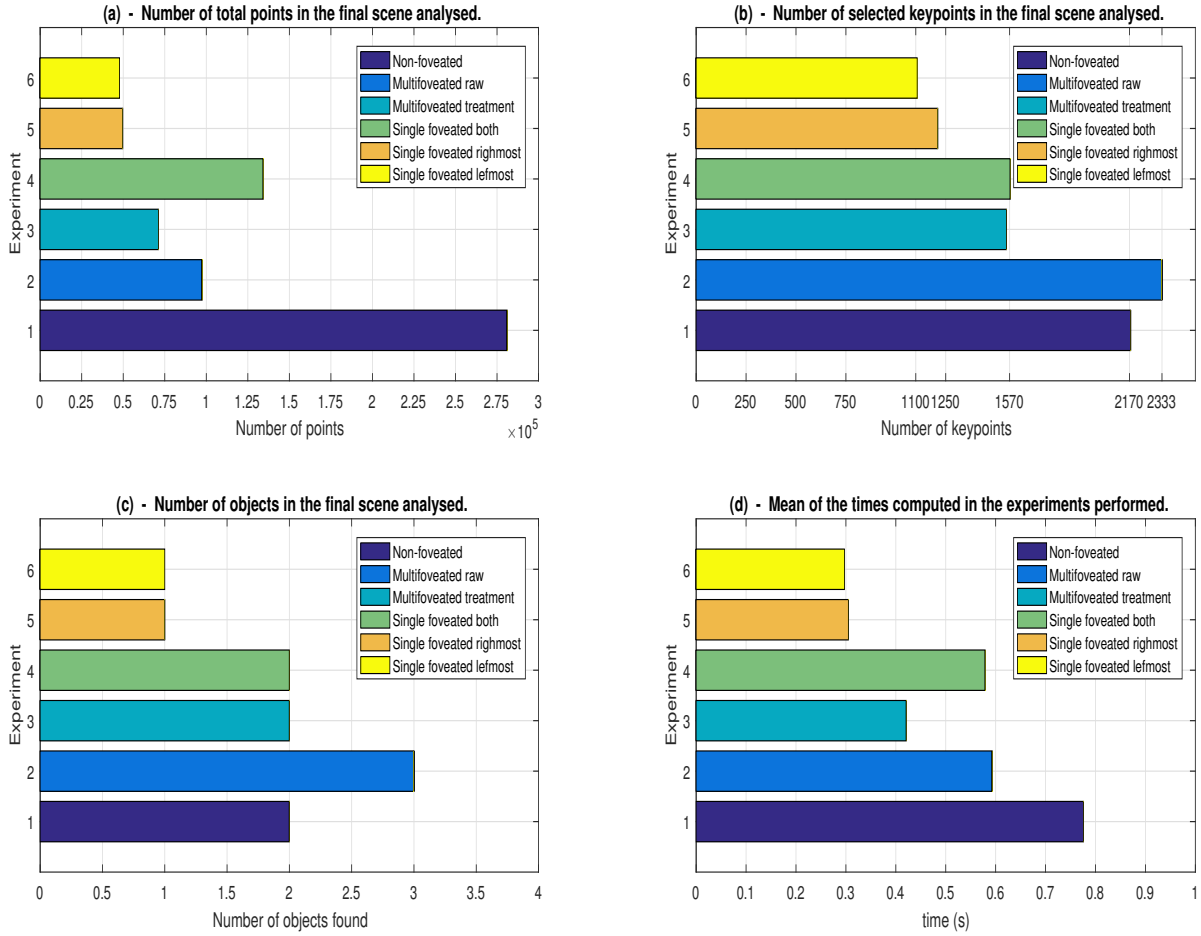

Fig. 2: The result group of the strategies performances carried out referring to the scene: (**a**) the total number of points in the final scene analyzed; (**b**) the result of the number of selected points of key-points in the final scene analyzed after execution of the strategies; (**c**) the amount of objects recognized for each configuration presented; and (**d**) the average of the times computed in each experiment performed where this result can be noticed in Table II (see more details in the text). The matches of the experiment numbers can be seen in Table I.

the identification of all objects, given the configurations already shown, is the multifoveated approach with redundancy treatment (our proposal). This is expected, since we have one of the largest points reductions relative to the original scene, as shown in Figure 2a. Table II presents the mean of the execution times in a more precise way, the maximum and minimum confidence intervals, the variances of the experiments and the reduction ratio in relation to our MFPC model to obtain a clearer analysis. Comparatively, the multifoveation with treatment (MFPC) has more than 25% of time reduction than the other strategies used. We consider that all objects should be identified in the scene.

Then, for the group of performance graphs (Figure 2), it is noted that the ability to reduce the number of points around the desired objects allows a reduction in execution time. However, it does not proportionately reduce the scene's descriptiveness, one of the positive points of the methods taken. It is also possible to observe that the amount of key-points is not necessarily a sign of improvement in the quantity and quality of the features as the multifoveated raw has more selected key-points than the original scene and, even then, there is the recognition of an object that does not exist in the scene. This fact occurred due to the redundancies generated, as already explained.

B. Results Analyzed from the Accuracy Group

Given these circumstances, we have drawn up charts to deal with the accuracy of the strategies (Figure 3). Figure 3a shows the matches that are made. That means the total number of selected points that could represent some point of the model object. It can be noticed that the approaches non-foveated, raw and FCAS are those that have a greater number of correspondences. The MFPC, FCRM and FCLM experiments have their respective expected match results, since they were more directed to the desired objects.

Analyzing the results by the number of valid hits, that is, the number of true-positive, we construct the graph represented in Figure 3b. An interesting fact in this analysis is the result obtained with the single fovea model that covers all the objects. It obtained the highest value of true-positive surpassing the non-foveated strategy. This fact can be explained by the configuration focus in the areas with high concentration of descriptiveness and similarity with the desired objects. We can

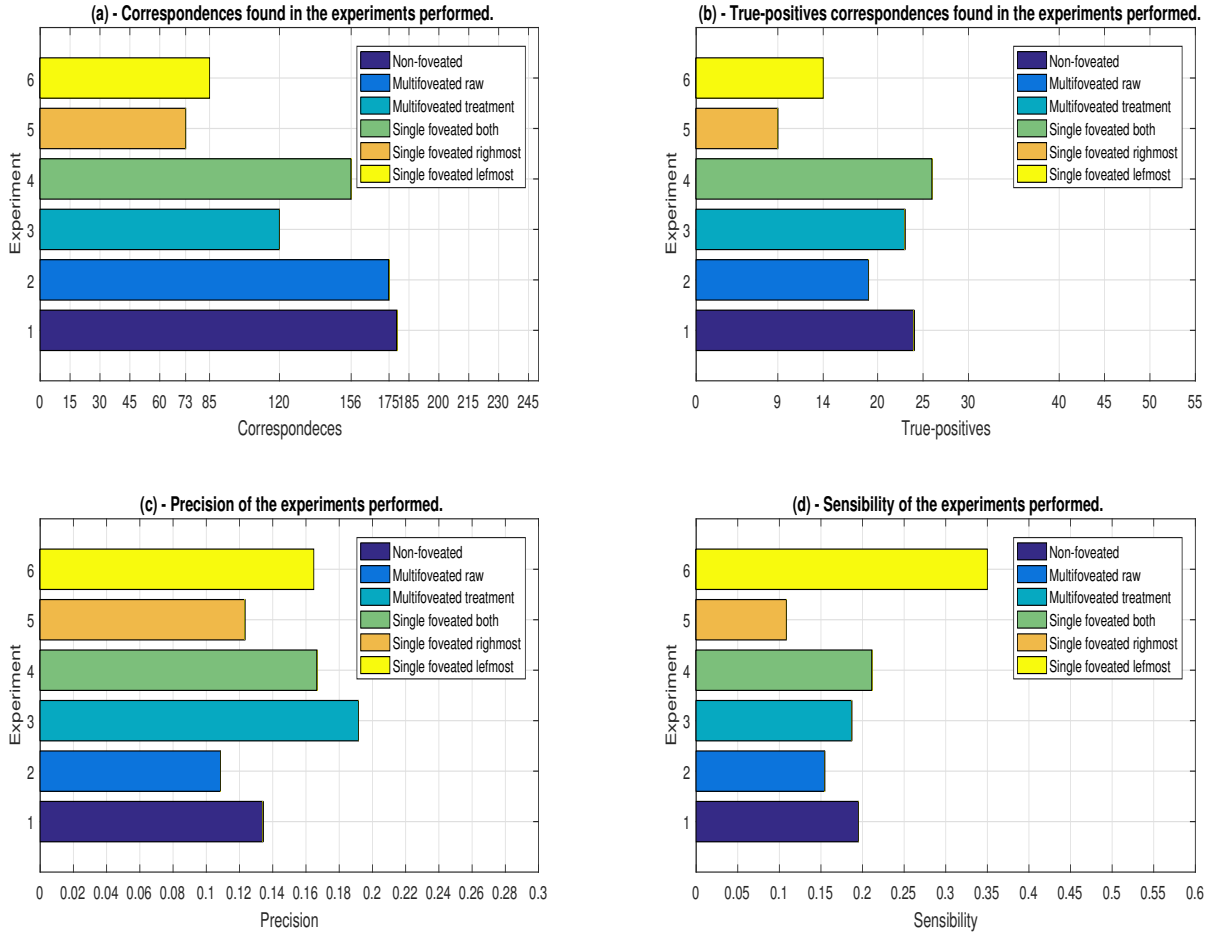

Fig. 3: The result group of the strategies accuracies carried out referring to the scene: (**a**) the total number of matches performed in the final scene analyzed; (**b**) the result of the number of true-positive selected in the scene; (**c**) the precision of each strategy; and (**d**) the sensitivity of each strategy (see more details in the text). The numbers of the experiments are mapped in Table I.

also notice that Experiments 1, 3, and 4 have similar results, as shown in Figure 3b. Practically, Experiments 1 (non-foveated) and 3 (MFPC) have the same true-positive numbers, showing that our proposal does not cause a large change in the validity of the data found, whereas the multifoveated raw model has a minimal advantage in the amount of valid matches compared to the Experiment 3, even getting much more matches and more points. It is observed that the duplication of points, which could be considered as reinforcement in the 3D environment, distorts the representation and the points' descriptiveness for the object recognition algorithm proposed by Tombari and Stefano [17].

The precision and sensitivity graphs are shown in Figure 3c,d, respectively. We observe that the foveated models usually maintain the precision and sensitivity as seen in the presented graphs. They can have superior performance of precision in relation to the model no-foveated, as seen in Figure 3c, which shows Experiment 3 with precision close to 20% and the no n-foveated approach (Experiment 1) with accuracy $< 15\%$. In relation to the sensitivity, it is the fraction of the relevant points identified. Given that the configuration for the extraction of key-points are the same for all strategies, we have the leftmost object with 40 in total to be identified, while the rightmost object has 83 key-points. This led to a higher sensitivity in the single fovea strategy that is positioned on the leftmost object, as shown in Figure 3d, since the difference is not so great among the true-positives in each strategy. Based on the presented sensitivity results, the foveated models have similar results in comparison to the original model (non-foveated), except for the multifoveated raw model, and the highest percentage is the FCAS (4) among the foveated approaches. Thus, for tunings and the scene chosen, we have seen that the proposed multifoveated brings improvements in the reduction of quantity of points, execution time and precision in relation to Strategies 1 and 4, in addition to noting that the multifoveated raw is not feasible for the various problems mentioned.

VI. Conclusions

This work has proposed an approach for reducing the amount of point cloud data captured by RGB-D sensors using multiple structures borrowed from the work of [1]. This mechanism has been integrated and tested in the object recognition task as exposed by Tombari and Stefano [17] and can be integrated into tasks involving visual attention control and recognition. The proposed mechanism provides a considerable reduction of

execution time in relation to other types of approaches between models and the original approach.

This article has three main contributions. The first one is the proposed mechanism that uses several foveated structures that can be found without producing redundant points in a PC. The second is that the work provides the conservation of the density hierarchy of the levels of the structures. The last contribution relies on the investigation of the problems related to FPC [1]. That investigation allowed noticing that the source of the problem at that scheme is caused by the descriptor used in the proposal, SHOT [18], which has sensitivity to the variation of the point density. This is solved with the improvements made by Salti [19] that greatly reduced the sensitivity, making it possible to perform the task without occurrence of detection problems.

PUBLICATIONS

Two articles have been published as results of the works done in this Master Thesis. In the first one, we helped on studying and implementing inverse methods for accelerating Visual SLAM. This initial work has been published as a conference paper at ICINCO 2017 and can be found at Souza et al. [2]. After this initial contribution on Visual SLAM, we came up that reducing the point cloud for later use in that approach was our goal, resulting in the MFPC approach above described. We changed from the VSLAM task to recognition, which is easier to test data reduction, being only this considered as the contribution in our Master thesis. A journal paper on this subject has been published at Sensors (A1 at Brazilian CAPES Qualis), which can be found at Oliveira et al. [16].

ACKNOWLEDGMENT

The authors would like to thank CAPES and CNPq for supporting this research.

REFERENCES

[1] R. B. Gomes, B. M. F. Silva, L. K. d. M. Rocha, R. V. Aroca, L. C. P. R. Velho, and L. M. G. Gonçalves, "Efficient 3d object recognition using foveated point clouds," *Compututer Graphhics*, vol. 37, no. 5, pp. 496–508, Aug 2013.

[2] A. Souza, L. Souto, F. F. de Oliveira, B. N. Datta, and L. M. G. Gonalves, "Using polynomial eigenvalue problem modeling to improve visual odometry for autonomous vehicles," in *Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics - Volume 2: ICINCO,*, INSTICC. SciTePress, 2017, pp. 502–507.

[3] B. M. F. da Silva and L. M. G. Gonçalves, "A fast feature tracking algorithm for visual odometry and mapping based on rgb-d sensors," in *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, Aug 2014, pp. 227–234.

[4] S. Lowry, N. Snderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, Feb 2016.

[5] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," *IEEE Transactions on Robotics*, vol. 30, no. 1, pp. 177–187, Feb 2014.

[6] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," in *CVPR 2011*, June 2011, pp. 441–448.

[7] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention," *The Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014, pMID: 24362813.

[8] L. M. Garcia, A. A. F. Oliveira, R. A. Grupen, D. S. Wheeler, and A. H. Fagg, "Tracing patterns and attention: humanoid robot cognition," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 4, pp. 70–77, Jul 2000.

[9] R. B. Gomes, L. M. G. Goncalves, and B. M. Carvalho, "Real time vision for robotics using a moving fovea approach with multi resolution," in *2008 IEEE International Conference on Robotics and Automation*, May 2008, pp. 2404–2409.

[10] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 243–254, Feb 2003.

[11] E.-C. Chang and C. K. Yap, "A wavelet approach to foveating images," in *Proceedings of the Thirteenth Annual Symposium on Computational Geometry*, ser. SCG '97. New York, NY, USA: ACM, 1997, pp. 397–399.

[12] A. Bernardino and J. Santos-Victor, *A Binocular Stereo Algorithm for Log-Polar Foveated Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 127–136.

[13] K. W. Chen, C. W. Lin, T. H. Chiu, M. Y. Y. Chen, and Y. P. Hung, "Multi-resolution design for large-scale and high-resolution monitoring," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1256–1268, Dec 2011.

[14] P. L. Camacho, F. J. Coslado Aristizabal, M. González García, and F. Sandoval Hernández, "Multifoveal imager for stereo applications," *International Journal of Imaging Systems and Technology, Vol. 12, No. 4*, pp. 149–165, 2002, diciembre 2002.

[15] J. A. F. Rodríguez, C. Urdiales García, A. J. Bandera Rubio, and F. Sandoval Hernández, "Nonuniform video coding by means of multifoveal geometries," *International Journal of Imaging Systems and Technology, Vol. 12, No. 1*, pp. 27–34, 2002.

[16] F. F. Oliveira, A. A. S. Souza, M. A. C. Fernandes, R. B. Gomes, and L. M. G. Goncalves, "Efficient 3d objects recognition using multifoveated point clouds," *Sensors*, vol. 18, no. 7, 2018.

[17] F. Tombari and L. D. Stefano, "Object recognition in 3d scenes with occlusions and clutter by hough voting," in *2010 Fourth Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, Nov 2010, pp. 349–355.

[18] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 356–369.

[19] S. Salti, F. Tombari, and L. Di Stefano, "Shot: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, Aug 2014.