ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY (ROMJIST)
Volume **25**, Numbers 3–4, 2022, 338–349

An RGB-D Descriptor for Object Classification

Erkut ARICAN* and Tarkan AYDIN

[1]Dept. of Computer Engineering, Bahcesehir University, Istanbul, Turkey
E-mails: erkut.arican@eng.bau.edu.t*, tarkan.aydin@eng.bau.edu.tr
* Corresponding author

Abstract. One of the main and active research areas in computer vision is the object detection which has various applications including image retrieval, video surveillance, robotics, etc. The main problem of object detection is, detecting instances of semantic objects of predefined classes (such as pedestrians, faces, or cars) in 2D images and videos. As 2D images of the objects include information about object appearance, most of the methods rely on pattern detection algorithms using appearance-based or feature-based techniques. Although the availability of 3D image data by using inexpensive depth cameras has made the problem more tractable, many researchers still tend to use similar concepts applied to the 2D instance problem. In this paper, we aim to develop a 3D descriptor that exploits the information in 3D data to address the many difficulties associated with object detection. This method adds depth information to Bag of Visual Words' feature extraction part which is a novel approach in the literature. The proposed 3D descriptor eliminates the disadvantages of brightness-based problems and improves the structure with depth information. This improvement gives better accuracy results compared to the original method providing a rational and useful method for 3D object detection.

Key-words: 3D descriptor; bag of visual words; computer vision; depth image; object detection; RGB-D.

1. Introduction

Object detection and classification which have been widely studied in literature, are very popular topics in computer vision, and used in a wide range of applications. Object detection means finding and identifying/classifying objects in images and videos of which methods can be classified as appearance-based and feature-based. Appearance-based methods such as skin color detection simply use color models such as RGB, HSI, TSL, and YCrCb, while feature-based methods exploit information in the shape-related brightness patterns of images. These methods are more popular because of their higher performance in terms of accuracy, and they include a

feature detection step, followed by a learning model that uses extracted local features to learn representations of target objects.

There are numerous feature detectors introduced in the literature. *Binary Robust invariant scalable keypoints* (BRISK) [1], Improved FAST [2], ORB [3], *Speeded Up Robust Feature* (SURF) [4], and *Scale Invariant Features Transform* (SIFT) [5] are the most frequently used feature detectors because of their robustness against geometric and photometric transformations. BRISK algorithm [1] combines a FAST-based detector and assembly of a bit-string descriptor. The paper introducing BRISK algorithm in the literature [1] indicates that although SIFT [5] and SURF [4] algorithms are widely used in the feature detection process. FAST [2] and BRIEF [6] algorithms offer real-time alternatives in the field. FAST [2] improves existing corner detectors to be faster and of high quality. However, there are some disadvantages such as being not robust to high-level noise and being dependent on a threshold. ORB [3] is based on BRIEF [6] and aims to solve the disadvantages of Improved FAST's [2] algorithm. It is robust to rotation, and noise and effective for real-time performance. H. Bay *et al.* [4] proposed the SURF algorithm which is a local feature detector and a descriptor. SURF is robust to noise, detection errors, and geometric and photometric deformations. On the other hand, SIFT performs reliable matching.

Since the number of detected features depends on the brightness and noise patterns in the images, it is impractical to standardize and limit the number of features for object types. This problem was addressed in *Bag of Visual Words* (BoVW) [7], a brightness dependence feature-based method, which creates a vocabulary and represents objects using a histogram of features to standardize representations. It creates a bag of keypoints and uses a multi-class classifier to determine the categories. Lazebnik *et al.* [8] suggest a robust and geometrical invariant structural representation by creating a hierarchical bag of words using the pyramid matching scheme [9]. However, the hierarchical bag of words also has some disadvantages such as being not suitable for heavy clutter, pose changes, and 3D image data. Utilizing the shape information may overcome brightness dependence and the hierarchical bag of words problems. Depth images are giving 3D information about the object's shapes and it can increase the effect of the structure. There are some studies to use BoVW with depth data [10–12] and a 3D descriptor is introduced to improve the bag of words method using RGB-D's depth information.

Many researchers have started to work on 3D object detection. In [13], the author detects 3D objects using multi-modalities. They use a combination of an image and dense depth map information. Since time consumption is very critical, this work is advantageous for time-saving on the training state and works in real-time. Their approach detects, with almost no false positive, 3D texture-less objects with heavy background clutter, illumination, and noise in real-time.

Johnson's paper [14] presents 3D shape-based object recognition and recognizes simultaneous multiple objects under clutter and occlusion. They construct 2D images and these images construct a local basis at an oriented point on the surface of an object. An oriented point is a 3D point with a surface normal. When they create images with 3D points, they can use 2D template matching. In addition, Sipiran's research paper [15] uses the Harris operator to find interest points detector for 3D objects. It has advantages such as being invariant to affine transformation, robust to noise, and different tessellations. In another work[16], the author propose a *3-dimensional SIFT* (3D SIFT) descriptor. They use video or 3D imagery and the Bag of words method. Sub-histograms are encoded in their 3D SIFT descriptor. Their study is very similar to SIFT [5]. They add time information, which is obtained from the video, to create a 3-dimensional SIFT descriptor. Furthermore, Mian's work [17] includes 3D model building and object recognition in which they use multiple unordered range images for creating 3D models which are

converted to multidimensional table representation and for recognition. Moreover, in [18], the author presents a global model description using oriented point pair features and matching the model locally with a fast voting scheme. The global model description is built in an offline phase and selecting a set of reference points from the scene is done online. On the other hand, study [19] uses the multi-value histogram and characterizes the local geometry. They use a machine learning algorithm for 3D geometric primitives. Their goal is to identify 3D points easily. In another paper, Rusu *et al.* [20] deal with the problem of aligning point cloud data views into a consistent global model. There is one more study by Rusu [21] trying to find a solution to the complexity problem for 3D registration for overlapping point cloud views. Apart from these, the research results reported in [22] address surface matching with local 3D descriptors in which they categorize available methods into two categories as signature and histograms. They explain the uniqueness and repeatability of the local reference frame. Besides, a sliding shapes method for 3D object detection which is proposed in [23], targets to solve major difficulties. In the sliding shapes method, they render a depth map with orientation, scale, 3D location, and camera tilt angle parameters. Sliding shapes combine point density feature, 3D shape feature, 3D normal feature, and TSDF feature; after that, it uses an SVM classifier for object detection. Another 3D object detection work created by Gupta [24] develops a new algorithm to find object boundary detection and hierarchical segmentation. Their system uses contour detection, segmentation, amodal completion, and semantic labeling.

This paper propose a system that combines *Bag of Visual Words* (BoVW) [7] with the SURF [4] and depth information to create a new 3D descriptor. BoVW [7] uses a base method with SURF [4] technique to extract feature vectors. This paper adds depth information while extracting feature vectors to create accurate 3D descriptors. Results show adding depth information has a huge impact on the accuracy rate compared to the original method.

The results given in this paper indicate that the proposed method of adding depth is analytical and practical for object classification, and it is a novel contribution to computer vision. This paper is organized as follows: explain the proposed method in Section 2, present the results in Section 3 and highlight the conclusions in Section 4.

2. Method

This section will give background information and explain the creation of a new 3D descriptor process.

2.1. Backgrond information

Bag of words is a document text classification method. It counts the words in a text and creates a vocabulary using a sparse histogram. In computer vision, the bag of words model uses local image features in its vocabulary Csurka's BoVW [7] and Sivic's paper [25] explain and use the bag of words model. In the present study, the BoVW method is used as a base method. Csurka [7] develops a descriptor for identifying an object. There are four main steps of BoVW:

- Determination of image patches,
- Creation of a vocabulary,
- Creation of bag of keypoints,
- Determination of categories for input images using a multi-class classifier.

It is easy to use and implement, however, it is view-dependent for 2D images, and structure information is lost.

Bilateral filters [26–28] express structure using color & 2D distance. The bilateral filter creates an average value for using nearby pixels with position and intensity in the sense that, the closer pixel and color belong to the same object. For instance, the texture on the object may have different colors so it is not an efficient way for object detection. Therefore, we use depth information in this study for much better performance. Adams [29] explained the generalized form for filtering which is changing values with a linear combination of other values. Using this definition, the following relation is proposed for generalized filtering:

$$\tilde{I} = \omega * I, \tag{1}$$

where Ĩ, I, and ω represent a filtered image, image, and weight value respectively. The structure can be better expressed using the 3D image.

Tomasi's study [28] explains bilateral filtering for gray and color images. This study shows a closeness function given by:

$$C(\eta, \rho) = e^{\left(-\frac{||\eta-\rho||^2}{2\sigma^2}\right)}, \tag{2}$$

where $C(\eta, \rho)$ is the geometric closeness, ρ is the center point of the neighborhood and η is the close point. Bilateral filtering is a good way for a grayscale or color image but it is inadequate for the shape of an object. Equations (1) and (2) give an insight to add depth value to solve the structure problem.

2.2. Creation of a new 3D descriptor

Adding depth information becomes a necessary step for eliminating the disadvantages of BoVW [7]. Using the four steps of BoVW specified above, this paper modifies and creates a new 3D descriptor using RGB-D's depth data. The use of depth information gives a huge impact on accuracy. The high rate of accuracy provides a better classification rate for the object. At the end of the process, we present a more reliable 3D descriptor to the literature.

Previous studies [10–12] provided the first insights for this study and the current proposed method consists of three parts which are 1) 2D image processing, 2) Depth image processing, and 3) Combining process. The method is explained step by step in Algorithm 1.

Algorithm 1 SURF + Depth Method

I. Read RGB image in grayscale
II. Detect SURF points from RGB image
III. For each SURF point selected as a center
  a. SFV ⟵ Extract SURF features from the RGB image
  b. Read depth image in grayscale
  c. Depth image ⟵ Apply Gaussian Filter
  d. Normalize the depth image between 0 and 1
  e. DFV ⟵ Extract features from depth image using SURF method
  f. Combine SFV with DFV using Equation (3)
  g. Return new descriptor

In the first part, the SURF method is used without estimating the orientation of the 2D image to extract keypoints. Examples of extracted SURF points with locations of interest are shown as circles in Fig. 1 using dataset images. The system uses these keypoints to create the *SURF feature vector* (SFV). The size of SFV is VP $\times$ 64, where VP is the number of valid SURF keypoints.

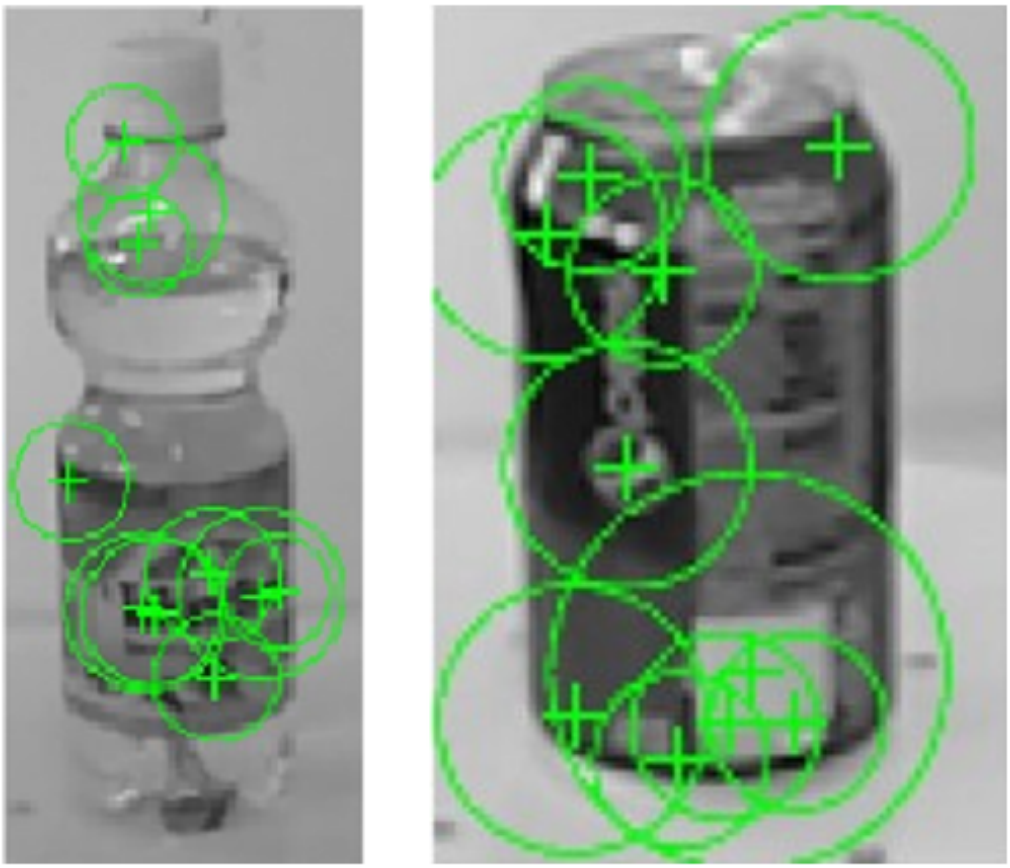

Fig. 1. SURF points example in water bottle and soda can.

In a depth image, an object's distance to a camera is very important. Depending on the distance of the object, depth matrix effect differs in the calculation. Thus, using relative distance value is important and the Gaussian filter is a crucial step in this current study. The Gaussian filter was applied to the depth image with $\sigma = 8$ and then normalized in the second part. This normalization gives us relative distance information. In this way, the object that is either far or close to the camera does not affect the proposed descriptor.

The next step was extracting the *depth image Feature Vector* (DFV) from depth images. Using the same keypoints which were extracted from a 2D image in the previous stage, the *depth image feature vector* (DFV) is extracted with the SURF method. The number of extracted SFV and DFV was the same and in the last step, SFV and DFV were combined to obtain a new 3D feature vector using element-wise multiplication:

$$NFV = SFV \odot e^{(DFV)}, \tag{3}$$

where NFV is the new feature vector and will give better accuracy results.

As mentioned in Section 2.1, original (or initial) BoVW methods use SURF for feature extraction, the K-Means clustering method for creating the vocabulary, and the multi-class SVM method for classification. In this study, after the NFV is produced, the K-Means clustering process creates the vocabulary. The number of clusters (K) is set to 20. Multi-class SVM with one-vs-all is used with some attributes. The kernel function is Polynomial Kernel and the Polynomial degree is 3. *Karush-Kuhn-Tucker* (KKT) threshold is selected as 10^{-2}. The iterative Single Data algorithm [30] is selected for multi-class SVM.

3. Results

RGB-D Object Dataset [31] is used in this paper. It is a Kinect Style dataset and it has 300 common household objects which are categorized into 51 categories. Examples of RGB images are presented in Fig. 2 and depth images are illustrated in Fig. 3.

Fig. 2. Dataset example: rgb image.

Fig. 3. Dataset example: depth image.

In this system, the SURF method was used to extract feature vectors. An approximate number of extracted features for each label in 10 label image sets is given in Table 1.

The proposed method was compared with the original BoVW method. BoVW method had two options for orientation of SURF feature vector. $BoVW_F$ represents the Upright options was *False*, which means need the image descriptors to capture rotation information. The other option is $BoVW_T$ and it represents Upright options was *True* which means estimating the orientation of the SURF feature vectors is not needed. BoVW and this method were run by 5 times with the same number of classes with random selection. The average accuracy result was taken and shown in Table 1. The proposed method gave better performance than the initial method in the different numbers of labels used in the dataset.

Table 1. Results of the methods

# of Class	$BoVW_F$	$BoVW_T$	PROPOSED METHOD
10	79.76%	76.25%	**89.42%**
20	65.04%	66.68%	**79.72%**
30	56.74%	59.50%	**66.18%**

Bo's study [32] uses extra dataset information and shows that RGB-D is giving better results than RGB. A comparison of some other studies' results presented in [32] is shared. Moreover, a few studies' results are shown in Table 2. Taking as a basis the average of the results presented in [32], it is noted that compared to the results obtained by using RGB, those obtained by using RGB-D are about 6% better for the category recognition and about 5% better for instances recognition.

Table 2. 10 label dataset detail comparison results

Categories	$BoVW_F$	$BoVW_T$	Proposed Method	# of Train Images	# of Test Images	# of Extracted Features
Apple	68.68%	52.20%	**76.81%**	425	182	4577
Ball	78.21%	61.79%	**89.27%**	574	246	3849
Bell Pepper	58.74%	55.68%	**76.74%**	444	190	8702
Binder	75.02%	73.43%	**96.43%**	497	213	34821
Coffee Mug	76.34%	**89.02%**	86.46%	384	164	15827
Keyboard	94.43%	95.23%	**99.54%**	410	176	139041
Pliers	94.24%	85.09%	**96.73%**	413	177	5274
Scissors	81.07%	75.51%	**90.37%**	438	187	11678
Soda Can	86.06%	**90.53%**	82.34%	440	188	14338
Water Bottle	84.76%	84.05%	**99.52%**	393	168	25010

The results in Table 2 together with the results presented in [32] show that using RGB-D images gives a better accuracy rate. A comparison of the accuracy rate values reported in other studies is given in Table 3.

Table 3. Accuracy rate values reported in other studies

	Instance Recognition			Category Recognition		
Methods	RGB	Depth	RGB-D	RGB	Depth	RGB-D
Linear SVM [31]	59.3%	32.3%	73.9%	74.3% ± 3.3	53.1% ± 1.7	81.9% ± 2.8
Unsupervised feature learning [32]	92.1%	51.7%	92.8%	82.4% ± 3.1	81.2% ± 2.3	87.5% ± 2.9

Table 1 shows more details for the 10 label dataset results. There are 4418 train and 1891 test images in this comparison. The bold number shows the best method accuracy result for that category. The proposed method gives better results in 8 categories.

Table 4 shows more details for the 20 label dataset results. There are 8655 train and 3704 test images in this comparison. The bold number shows the best method accuracy result for that category. The method gives better results in 16 categories.

Table 4. 20 label dataset detail comparison results

Categories	$BoVW_F$	$BoVW_T$	Proposed Method	# of Train Images	# of Test Images
Apple	40.66%	38.13%	**68.79%**	425	182
Ball	26.43%	35.36%	**78.13%**	574	246
Bell Pepper	22.00%	25.58%	**60.21%**	444	190
Binder	62.63%	65.63%	**89.95%**	497	213
Bowl	66.32%	68.39%	**98.74%**	406	174
Calculator	84.48%	84.71%	**86.09%**	406	174
Camera	57.65%	**59.14%**	36.79%	436	187
Cap	87.55%	86.70%	**99.04%**	440	188
Cell Phone	57.55%	68.47%	**73.37%**	381	163
Cereal	84.56%	92.28%	**97.54%**	399	171
Coffee Mug	57.93%	**73.54%**	67.56%	384	164
Comb	**68.05%**	66.86%	62.01%	396	169
Dry Battery	56.47%	41.77%	**77.53%**	397	170
Flashlight	65.53%	64.09%	**66.19%**	424	181
Food Bag	97.20%	98.56%	**99.32%**	552	236
Keyboard	88.52%	88.41%	**98.86%**	410	176
Pliers	85.65%	69.15%	**94.80%**	413	177
Scissors	59.79%	56.90%	**80.86%**	438	187
Soda Can	56.38%	**76.70%**	61.70%	440	188
Water Bottle	75.48%	73.10%	**96.79%**	393	168

Table 5 shows more details for the 30 label dataset results. There are 14117 train and 6044 test images in this comparison. The proposed method gives better results in 16 categories which are apple, ball, bell pepper, binder, bowl, calculator, cap, cell phone, cereal, dry battery, glue stick, hand towel, keyboard, pliers, scissors, and water bottle. For 'food can & soda can' and 'food bag & instant noodle' categories, the proposed method gives close but not good enough results because of the similarity of objects and depth images.

Table 5. 30 label dataset detail comparison results

Categories	$BoVW_F$	$BoVW_T$	Proposed Method	# of Train Images	# of Test Images
Apple	34.28%	11.10%	**61.87%**	425	182
Ball	24.39%	23.09%	**59.11%**	574	246
Bell Pepper	16.95%	13.69%	**48.53%**	444	190
Binder	43.66%	64.13%	**84.23%**	497	213
Bowl	62.53%	64.37%	**96.90%**	406	174
Calculator	75.29%	74.25%	**75.40%**	406	174
Camera	50.16%	**56.26%**	29.20%	436	187
Cap	85.21%	82.66%	**97.34%**	440	188
Cell Phone	52.88%	60.61%	**65.03%**	381	163
Cereal	75.79%	85.96%	**90.53%**	399	171
Coffee Mug	54.51%	**72.56%**	54.51%	384	164
Comb	**69.11%**	53.73%	46.39%	396	169
Dry Battery	08.47%	23.88%	**35.06%**	397	170
Flashlight	**66.63%**	58.67%	60.22%	424	181
Food Bag	95.85%	**96.53%**	93.47%	552	236
Food Box	58.90%	**70.00%**	56.69%	551	236
Food Can	40.00%	**63.08%**	62.11%	575	247
Food Cup	51.53%	**59.58%**	39.75%	550	236
Food Jar	72.17%	**76.69%**	56.61%	538	230
Garlic	**34.53%**	31.45%	29.40%	546	234
Glue Stick	16.75%	25.67%	**48.17%**	561	240
Greens	72.75%	**82.84%**	66.97%	510	218
Hand Towel	37.19%	26.47%	**84.34%**	548	235
Instant Noodle	81.21%	**81.56%**	67.01%	540	231
Keyboard	83.75%	86.02%	**93.30%**	410	176
Kleenex	80.69%	**82.92%**	79.83%	543	233
Pliers	84.75%	72.43%	**92.43%**	413	177
Scissors	54.97%	54.22%	**75.40%**	438	187
Soda Can	55.00%	**70.85%**	43.08%	440	188
Water Bottle	62.14%	59.76%	**92.74%**	393	168

4. Conclusions

Object detection is a very popular topic in computer vision, and there are various studies available in the literature. Through the breakthrough in technology and research, community can easily access 3D data, so 3D descriptor becomes an important topic. In this study, the system

combines BoVW with RGB-D's depth information to create a new 3D descriptor. This method eliminates the disadvantages of BoVW, and experiments show us that the proposed novel method gives a better accuracy rate compared to the original BoVW method. As a result, the proposed 3D descriptor gives a good performance with Kinect-style 3D datasets for 3D object detection.

Acknowledgment. This study is a part of Bahcesehir University Doctoral Programme's PhD Dissertation.

References

[1] S. LEUTENEGGER, M. CHLI and R. Y. SIEGWART, *BRISK: Binary Robust Invariant Scalable Keypoints*, Proceedings of 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 2548–2555, 2011.

[2] E. ROSTEN AND T. DRUMMOND, *Machine learning for high-speed corner detection*, in European Conference on Computer Vision, Springer, Berlin, Heidelberg, pp. 430–443, 2006.

[3] E. RUBLEE, V. RABAUD, K. KONOLIGE and G. BRADSKI, *ORB: An efficient alternative to SIFT or SURF*, Proceedings of 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 2564–2571, 2011.

[4] H. BAY, T. TUYTELAARS, and L. VAN GOOL, *SURF: Speeded Up Robust Features*, in European Conference on Computer Vision, Springer, Berlin, Heidelberg, pp. 404–417, 2006.

[5] D. G. LOWE, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision **60**, pp. 91–110, 2004.

[6] M. CALONDER, V. LEPETIT, C. STRECHA and P. FUA, *BRIEF: Binary Robust Independent Elementary Features*, in: European Conference on Computer Vision, Springer, Berlin, Heidelberg, pp. 778–792, 2010.

[7] G. CSURKA, C. DANCE, L. FAN, J. WILLAMOWSKI and C. BRAY, *Visual categorization with bag of keypoints*, Proceedings of 2004 Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, pp. 1–2, 2004.

[8] S. LAZEBNIK, C. SCHMID and J. PONCE, *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, pp. 2169–2178, 2006.

[9] K. GRAUMAN and T. DARRELL, *The pyramid match kernel: Discriminative classification with sets of image features*, Proceedings of 10th IEEE International Conference on Computer Vision, Beijing, China **1**, pp. 1458–1465, 2005.

[10] E. ARICAN and T. AYDIN, *Object Detection With RGB-D Data Using Depth Oriented Gradients*, Book of Proceedings-International Conference on Engineering and Natural Sciences, Budapest, Hungary, 2017.

[11] E. ARICAN and T. AYDIN, *A new descriptor for 3D object detection using RGB-D*, Proceedings of 2018 International Conference on Science and Technology, Prizrrn, Kosovo, pp. 1–6, 2018.

[12] E. ARICAN and T. AYDIN, *3D object detection using a new descriptor with RGB-D*, Bilge International Journal of Science and Technology Research **3**(1), pp. 58–62, 2019.

[13] S. HINTERSTOISSER, S. HOLZER, C. CAGNIART, S. ILIC, K. KONOLIGE, N. NAVAB and V. LEPETIT, *Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes*, Proceedings of 2011 International Conference on Computer Vision, Barcelona, Spain, pp. 858–865, 2011.

[14] A. E. JOHNSON and M. HEBERT, *Using spin images for efficient object recognition in cluttered 3D scenes*, IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(5), pp. 433–449, 1999.

[15] I. SIPIRAN and B. BUSTOS, *A Robust 3D interest points detector based on Harris operator*, Proceedings of 3DOR@ Eurographics, Norrköping, Sweden, pp. 7–14, 2010.

[16] P. SCOVANNER, S. ALI and M. SHAH, *A 3-dimensional sift descriptor and its application to action recognition*, Proceedings of 15th ACM International Conference on Multimedia, New York, NY, USA, pp. 357–360, 2007.

[17] A. S. MIAN, M. BENNAMOUN and R. OWENS, *Three-dimensional model-based object recognition and segmentation in cluttered scenes*, IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(10), pp. 1584–1601, 2006.

[18] B. DROST, M. ULRICH, N. NAVAB and S. ILIC, *Model globally, match locally: Efficient and robust 3D object recognition*, Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, pp. 998–1005, 2010.

[19] R. B. RUSU, Z. C. MARTON, N. BLODOW and M. BEETZ, *Learning informative point classes for the acquisition of object model maps*, Proceedings of 10th International Conference on Control, Automation, Robotics and Vision, Hanoi, Vietnam, pp. 643–650, 2008.

[20] R. B. RUSU, N. BLODOW, Z. C. MARTON and M. BEETZ, *Aligning point cloud views using persistent feature histograms*, Proceedings of 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, pp. 3384–3391, 2008.

[21] R. B. RUSU, N. BLODOW and M. BEETZ, *Fast Point Feature Histograms (FPFH) for 3D registration*, Proceedings of 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, pp. 3212–3217, 2009.

[22] F. TOMBARI, S. SALTI and L. DI STEFANO, *Unique signatures of histograms for local surface description*, in European Conference on Computer Vision, Springer, Berlin, Heidelberg, pp. 356–369, 2010.

[23] S. SONG and J. XIAO, *Sliding shapes for 3D object detection in depth images*, in European Conference on Computer Vision, Springer, Cham, pp. 634–651, 2014.

[24] S. GUPTA, P. ARBELEZ and J. MALIK, *Perceptual organization and recognition of indoor scenes from RGB-D images*, Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, pp. 564–571, 2013.

[25] J. SIVIC and A. ZISSERMAN, *Video Google: a text retrieval approach to object matching in videos*, Proceedings of 9th IEEE International Conference on Computer Vision, Nice, France, pp. 1470–1477, 2003.

[26] V. AURICH and J. WEULE, *Non-linear gaussian filters performing edge preserving diffusion*, in Mustererkennung 1995, Springer, Berlin, Heidelberg, pp. 538-545, 1995.

[27] S. M. SMITH and J. M. BRADY, *SUSAN–A New Approach to Low Level Image Processing*, International Journal of Computer Vision **23**, pp. 45–78, 1997.

[28] C. TOMASI and R. MANDUCHI, *Bilateral filtering for gray and color images*, Proceedings of 6th International Conference on Computer Vision, Bombay, India, pp. 839–846, 1998.

[29] A. ADAMS, N. GELFAND, J. DOLSON and M. LEVOY, *Gaussian KD-trees for fast high-dimensional filtering*, in ACM SIGGRAPH 2009 papers (SIGGRAPH '09), Association for Computing Machinery, New York, NY, pp. 1–12, 2009.

[30] V. KECMAN, T.-M. HUANG, M. VOGT, *Iterative single data algorithm for training kernel machines from huge data sets: Theory and performance*, in Support Vector Machines: Theory and Applications, Springer, Berlin, Heidelberg, pp. 255–274, 2005.

[31] K. LAI, L. BO, X. REN and D. FOX, *A large-scale hierarchical multi-view RGB-D object dataset*, Proceedings of 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, pp. 1817–1824, 2011.

[32] L. BO, X. REN and D. FOX, *Unsupervised feature learning for RGB-D based object recognition*, in Experimental Robotics, Springer, Heidelberg, pp. 387–402, 2013.