Indonesian Journal of Electrical Engineering and Computer Science
Vol. 23, No. 2, August 2021, pp. 1073∼1083
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v23.i2.pp1073-1083

Hybrid features for object detection in RGB-D scenes

Sari Awwad[1], Bashar Igried[2], Mohammad Wedyan[3], Mohammad Alshira'H[4]
[1,2]Department of Computer Science and Applications, The Hashemite University, Zarqa, Jordan
[3]Autonomous Systems Department, Al-Balqa Applied University, Salt, Jordan
[4]Department of Information Systems, Al alBayt University, Mafraq, Jordan

Article Info

Article history:

Received May 28, 2021
Revised Jul 6, 2021
Accepted Jul 14, 2021

Keywords:

Binary SVM
LDBD
Object detection
RGB-D object dataset
SIFT
Sliding window

ABSTRACT

Object detection is considered a hot research topic in applications of artificial intelligence and computer vision. Historically, object detection was widely used in various fields like surveillance, fine-grained activities and robotics. All studies focus on improving accuracy for object detection using images, whether indoor or outdoor scenes. Therefore, this paper took a shot by improving the doable features extraction and proposing crossed sliding window approach using exiting classifiers for object detection. In this paper, the contribution includes two parts: First, improving local depth pattern feature alongside SIFT, and the second part explains a new technique presented by proposing crossed sliding window approach using two different types of images (colored and depth). Two types of features local depth patterns for detection (LDPD) and scale-invariant feature transform (SIFT) were merged as one feature vector. The RGB-D object dataset has been used and it consists of 300 different objects, and includes thousands of scenes. The proposed approach achieved high results comparing to other features or separated features that are used in this paper. All experiments and comparatives were applied on the same dataset for the same objective. Experimental results report a high accuracy in terms of detection rate, recall, precision, and F1 score in RGB-D scenes.

Corresponding Author:

Sari Awwad
Department of Computer Science and Applications
The Hashemite University, Zarqa, Jordan
Email: sari@hu.edu.jo

1. INTRODUCTION AND BACKGROUND

Object detection has been widely used in different computer vision applications to imitate human visual perception in finding the most important object(s) in a scene. The advancement of technology in imaging machines and the growing need for applications in different fields, whether for monitoring and tracking or medical purposes. In artificial intelligence applications, object detection aims to extract detailed information and features about the various objects in a scene. Generally, this information are performed over RGB or grey-level images are acquiring from imaging machines [1].

In the last two decades, later propels in camera innovation have led to the advancement of reasonable imaging devices that can capture depth images alongside common RGB. These types of cameras give rich information about distances between each pixel in an object and depth sensor. Microsoft Kinect, Asus Xtion, and Stereo labs ZED are examples of cameras that contains depth sensors [2]. Therefore, scientists reconstruct their researches and concentrate on this type of research by using depth information that is given by these type of cameras.

In object detection area, there are different angles of challenges are facing researchers such as scenes type, window size for detected object, learning methods, processing time complexity, boosting schemes, and features selection [3]. Ideally, feature descriptors provide detailed information of an object and its variations about the background regardless the size of object. Moreover, the provided information by descriptors should be rich in order to enable robust detection or localized of that object [4], [5]. Mainly, there are two main types of feature descriptors: i) global descriptor, ii) local descriptor, where the first one is concerned with visual features as a whole image, while the second one is concerned with a description of patch or part of an image which are specific for localized considerations [6], [7].

Global descriptors are concerned with top-down approaches, where the primary stage of this approach comprises of the localization an object inside the image utilizing either tracking or foreground detection using background subtraction. Regularly, global descriptors are more touchy with changes in region of interests (ROI), noises of background and occlusions. Generally, when the space permits for solid levels of control over these variables, global descriptors ordinarily perform well [7], [8].

Oppositely, local descriptors are relevant to local patterns that is concerned with a specific area of an image, which makes this type of descriptors less affected by changes of appearance, background clutters, or occlusions between foreground objects. Local descriptors are utilized to characterize and describe interested points in groups of an independent patches. This type of descriptors is more reasonable for still images than global descriptors [6].

Depth images contain the distances between each pixel in acquired object and the sensor of the camera, where these distances are calculated relatively to the position of the camera. Depth information is considered as one of the important tasks for a computer vision system [9]. Depth features achieves relatively good performance to add certain characteristics to depth map sequences, these depth maps are very effective to solve some challenges in object detection or tracking, where the distance value will be the solution in objects occlusion [10]. Moreover, depth images and its information play a big role in artificial intelligent applications and computer vision, whether at the level of traceability or object detection specially in areas that are sensitive to privacy, such as hospitals and nursing homes. Figure 1 is an example of depth images alongside RGB that are used in our experiments of RGB-D object dataset.

Figure 1. Depth images: an example of depth images alongside RGB [11], [12]

Paper structure consists of sections as follows; section 1 introduction, provides a rich information and a high-level overview of the context regarding object detection and outcomes in which this research is being completed. Section 2 describe literature review, background on approaches, techniques, classifiers, and features that were existed for object detection. Sections 3 and 4 explains methodology and experiments, displaying our approach and discussing experiments and results and last section 5 conclusion and future work.

The process of selecting suitable features plays an important role in developing object detection and increasing the accuracy, where the feature descriptors of the target object give all characteristics and measurements within an object surrounded by a window. Moreover, features feed the training model with all necessary information to perform the aimed tasks, whether the task is recognition, detection, or identification [13], [14].

In this area, most researchers seek to develop the results of object detection by developing the existing techniques in terms of using classifiers that have achieved good results in other research topics in computer visions such as tracking and activity or object recognition. While others have pre-processed images to develop results or by modifying the sequence of steps in their methodologies for the same purpose [15].

Therefore, the main problem is how to improve and increase the detection rate by utilizing depth images alongside RGB. The contribution of this paper is proposing a new technique, where it is consisted of the

following parts: the first one using cross sliding window approach, followed by improving local depth patterns features, finally build a combinational feature vector by utilizing depth images alongside RGB. Moreover, This paper has aimed to answer the following question: does selecting and combining features make a difference in developing the results for object detection?.

2. LITERATURE REVIEW

This section browsed recent researches that are focused in its study on classifiers, features, and type of dataset that used for object detection. Numerous strategies are introduced spread over various fields of research. Researches related to object detection addresses various issues depending on the type of applications associated with the research [16], [17].

There are different Types of models or classifiers that are used for object detection. For instance Researchers in [18], [19] used deep learning classifier for object detection in robotics domain, and they applied their experiments on RGBD using a specific camera for this purpose. Others used boost classifier by using a one-class universal detector to repeatedly transfer information from the source domain to the target domain and learn the target-domain detector. The target-domain detector improves the one-class universal detector by mining box-level pseudo ground facts in each iteration [20]. Yan *et al.* [21] designed a deep learning network for object detection based on merging the geometric data (3D) and texture data of two-dimensional (2D). To solve the issue of one sensor, they used an inverse mapping level and a gathering level to merge the one or more input of RGB datum with the geometric input of point cloud data and designed a top gathering layer to transact with the data of multiple vision cameras. Also, to resolve the fault of the procedure to detect the 3D object founded on the area proposal network procedure, they used the Hough-voting procedure performed by a deep neural network to detect objects. Experimental outcomes showed that their algorithm had a 1.06% decreased in average accuracy contrast to PointRCNN in simple vehicle object detection, however, their method demanded 37.7% less time to compute than PointRCNN under the environment of the same equipment. Also, their method improved the average accuracy by 1.14% contrast to PointRCNN in hard vehicle item detection.

But regarding the dataset type that used for object detection, over the last few years, a number of datasets have been acquired . For instance, in [22] coloured and depth images were used their proposed approach, it is acquired using mobile-manipulator in areal world environment, also Cheng *et al.* [23] used a dataset consists of 135 RGB-D images that were acquired by Kinect device, three persons had been asked to label the object in each image.

Deductively, the type of dataset differs from one research to another based on the purpose of the research, whether to improve results or to a specific application, for example in [24] used object detection for surveillance purposes, they used specific objects classification in airports, such as people, bags, trolleys. While in [21], [25], [26], the authors studied 3D object detection using RGB-D data scenes in outdoor [27] and indoor [28]-[30], or both. some of selected related studies will be described in details in the following.

From another perspective, different approaches and were proposed for object detection in RGB-D. For instance, [26], [28], the raw point clouds were directly operated by popping up RGB-D scans. The preciseness that localizes objects in point clouds of big scale scenes was the main challenge of this method. Rather than solely based on 3D proposals, they proposed a new method that invested both promoted 3D deep learning for object localization and mature 2D object detectors, leading efficiency in addition to rising recall for even tiny objects. leverages from learning straightway in raw point clouds, the proposed method was as well as able to estimate 3D bounding boxes in a precise way even in huge occlusion or with extremely sparse points, the proposed method outperformed the latest technology by distinguishing margins and having the capability in the real-time.

Also, Authors in [29], [30] presented a mechanism that makes 3D bounding borders surrounding items in an RGB-D sight. their technique made best used of the 2D data to rapidly decrease the exploring area in 3D, investing in the latest 2D objects technology recognition approaches. Then they used the 3D data to place, orient, and mark surrounding borders about items. They estimated the orientation for every item independently way, by using the prior mechanisms that use simple data. item positions and volumes in 3D were learned utilizing a multilevel perceptron. Finally, they refine their recognizes relied on objects group relations within a scene. When evaluated the recognition techniques that operated nearly completely in the scattered 3D space, wide assessments on the" SUN RGB-D" dataset showed that their proposed technique was much quicker (4.1 seconds/image) in recognizing 3-D items in RGB-D pictures and process better than the new technology that

was slower and comparing to the technique that was two times of magnitude slower. This research hinted at the idea that 2D-driven item recognizing in 3D.

On other hand, in [31] used depth and colour data to determine automatically the location of an object and minimized the difficulty of visual analysis by using salient item recognition for RGB-D pictures. They proposed salient item recognition by convolution neural network with a single stream. this proposed technique was done as following, to produce multiple rank features that presented the most main feature for RGB-D the first RGBD 4 inputs was provided into VGG-16 net to the image. Although the salient objects can detect and localize by the coarse saliency chart of the deepest features, absence of the borders and subtle compositions. The current research has focused on object detection using hybrid features (LDPD + SIFT) with crossing sliding window and using binary SVM as classifier.

3. THE PROPOSED APPROACH

This section browses main stages and explains its steps in details of our approach, where the proposed approach is divided into two parts. The first one is training stage that includes: create bounding box around the ground truth manually, features extraction (local depth patterns and SIFT features), build feature vectors, and build training model for each object using binary SVM (support vector machine). While the second part starts with features extraction from each bounding box that is created by cross window approach, then create feature vector for testing using binary SVM, this process will be repeated until object detected. Figures 2 and 3 show the steps of main stages (training and testing) that are used in the proposed approach.

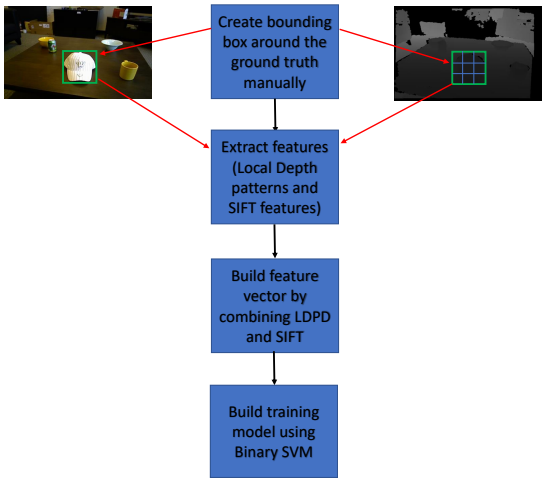

Figure 2. Training model stage

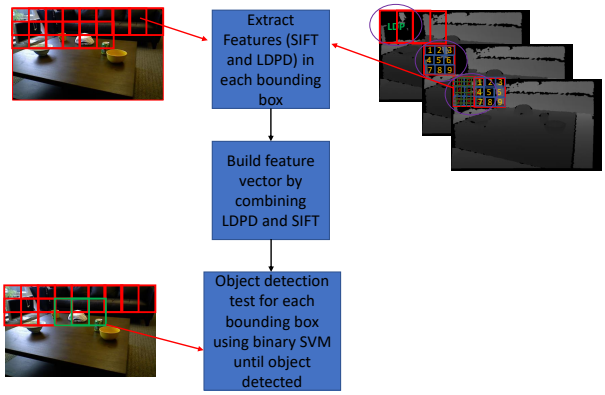

Figure 3. Testing stage

3.1. Features extraction

In this section, SIFT and LDPD features will be presented as main step in each stage, whether it was training stage or testing stage.

3.1.1. SIFT features

Many local features are available that used to detect interesting points on the object, where these points will be used to provide a 'feature' to consist a descriptor of the object. These descriptors can be used for various purposes in computer vision and artificial applications like object recognition, detection, and tracking. Many procedures can be considered when extracting local features and how to record these features. SIFT provide feature description of an object, where these features are not affected by many of the complications experienced in other methods, such as image rotation and scaling. After important points detected, the features are generated and transformed into features vectors. Figure 4 is an example of local descriptor [32].

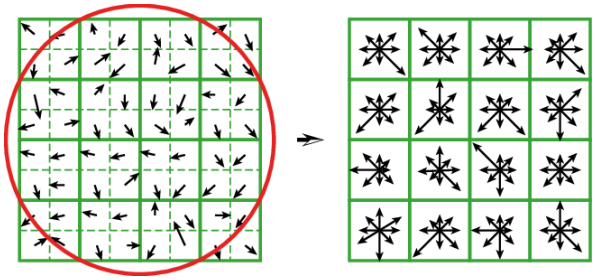

Figure 4. Local descriptor: an example of SIFT descriptor computations [32]

In this paper, local SIFT descriptor has been used for the purpose of object detection. In the proposed approach, each image has been divided into bounding boxes, where each bounding box has SIFT descriptor as seen in the Figure 5. Each feature vector size is 128. The size of each bounding box is variant based on the object, and the X,Y coordinates that are given by the RGB-D object dataset.

3.1.2. Local depth patterns for detection (LDPD)

In this paper, LDP feature has been improved and used alongside SIFT for its high effectiveness in object tracking and fined-activity recognition [33], [34]. For this reason, authors have decided to improve LDP feature for object detection to be called (LDPD) alongside SIFT features. In the proposed approach, each image has been divided into cross-bounding boxes, where each bounding box crosses the other one from half. Each bounding box has been divided into grid of depth descriptor, and each descriptor has been divided into small patches.

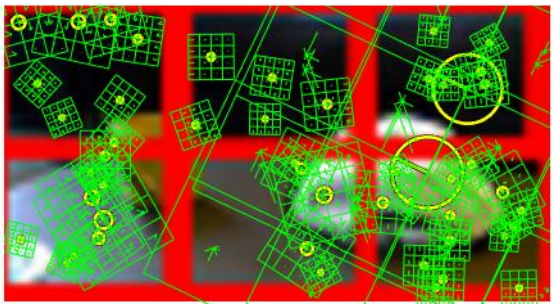

Figure 5. Local descriptor: an example of SIFT features in each bounding box

Therefore, for our approach feature (named LDP for detection, or LDPD as abbreviation). As mentioned before, Each bounding box has been divided into grid of LDPs named $HORD$ (Horizontal Descriptor) and $VERD$ (Vertical Descriptor), where the values of $HORD$ and $VERD$ are 3 and 3 respectively. That means each bounding box has 9 LDPs, and each LDP was divided into 3×3 cells. Each LDP value was computed by concatenating the differences between the depth average between adjacent cells in a patch. Figure 6 displayed an example how was the LDPD constructed. The (1) displayed the size of each LDPD vector.

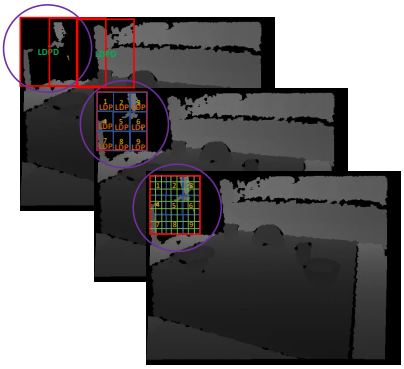

Figure 6. LDPD: an example of LDPD feature construction

Based on (1), the dimension of LDPD in the proposed approach is 324 dimensions. All detailed steps for LDPD computations are displayed in Algorithm 1.

$$size(LDPD) = HORD \times VERD \times \binom{3*3}{2} \tag{1}$$

3.2. Binary support vector machine

In this research, object detection has been considered as a binary classification problem, while the sequence of crossing bounding boxes are generated to detect the targeted object, each bounding box was tested using binary classifier to give us -1 or +1 label, where -1 is a wrong detection and +1 is a true detection. For this purpose, binary SVM has been used in the proposed approach.

Binary support vector machine was proposed by Vapnik [35] to find an optimal classification for two-label classes problem. Binary SVM classifier working on finding an optimal hyperplane between two classes, and that will be done by using a part of dataset as training data. The training objective is displayed in (2). SVM score is express by $w^T x + b$, where the w is a weight vector and b is the bias. The hyperplane in x-space is defined by $w^T x + b = y$ and the hyperplane that is separating between two classes expressed by $w^T x + b = 0$. Figure 7 is an example showing how the binary SVM working in the proposed approach whether in training stage or testing stage.

Algorithm 1: Local depth patterns algorithm for object detection

Input: Bounding window
Output: LDPD
// creates LDPD without value
$LDPD = \emptyset$
for $ro = 1$ *to* $VERD$ **do**
  for $co = 1$ *to* $HORD$ **do**
    // creates lcal descriptor without value
    $LDP(ro, co) = \emptyset$
    for $hor = 1$ *to* 9 **do**
      for $ver = hor + 1$ *to* 9 **do**
        // computes the subtraction between adjacent cells
        $sub(hor, ver) = |averagedepth(hor) - averagedepth(ver)|$
        $LDP(ro, co) = merge(LDP(ro, co), sub(hor, ver))$
    $LDPD = merge(LDP(ro, co))$

$$\begin{aligned} w^*, b^* &= \underset{w,b,\xi \geq 0}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \\ s.t.\ \ & y_i(w^T x_i + b) \geq 1 - \xi \quad i = 1 \ldots N \end{aligned} \tag{2}$$

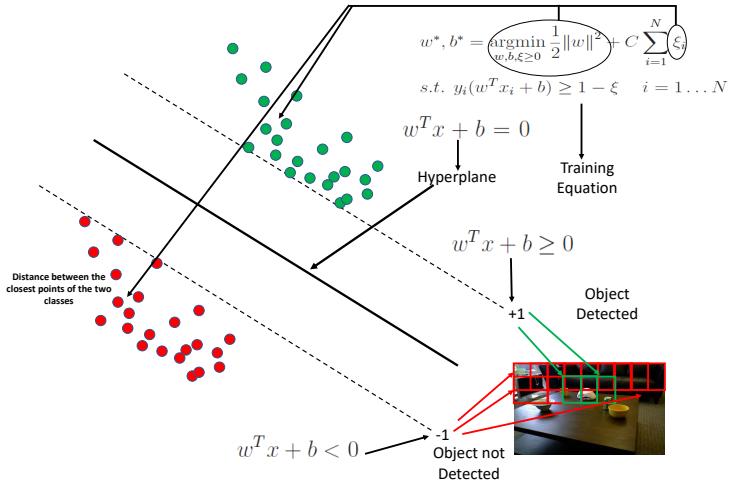

Figure 7. Binary SVM: showing the details of using binary SVM in the propose approach

4. EXPERIMENTS

Experiments section browses and describes dataset that was used in the experiments, evaluation part that was used in the proposed approach, and discusses the experimental results in terms of detection ate, recall, precision, and F1 score.

4.1. Dataset

RGB-D Object Dataset [11] has been used in the proposed approach, where this dataset contains 300 objects that are distributed in 51 categories, this dataset was acquired by using Kinect camera. The main advantage of the dataset that contains on hundreds of RGB and depth scenes for each object, where an object is available in each scene but in different: location, angle,colour, distance from camera sensor.

In this paper, seven different objects have been tested and used for object detection (bowls, caps, food plate, bell pepper, cereal boxes, coffee mugs, and soda cans), where these objects are a subset of the objects in the RGB-D Object Dataset under title ”RGB-D Scenes Dataset v.2/rgbd scenes v2 imgs.zip”. Figure 8 is a sample of various scenes that were used in the proposed approach.

Figure 8. Dataset: sample of various scenes that were used in the proposed approach

4.2. Evaluation

In the proposed approach, the evaluation process has been achieved based on the object is detected or not as binary classification problem. Therefore, separated training model has been built for each object by creating manual bounding box around the ground truth object based on X,Y coordinates that are available in the RGB-D object dataset. This process has been performed by using n-fold validation technique, where the group of scenes has been divided into five sets, in each iteration one of them for testing and the remaining for training.

In the testing stage, sliding of crossing windows are generated, where each window is tested if it contains the ground truth object or not by using binary SVM as mentioned before. Three main terms (Recall, Precision, and F1 score) have been measured in the evaluation process by using true positive (TP) , false positive (FP), true negative (TN), and false negative(FN). TP means correct match detected object with ground truth, but FP means wrong match detected object with ground truth, while the TN means correct undetected object with current location, but the FN means wrong undetected object with current location, where the current location is ground truth. The (3), (4), and (5) displayed how recall, precision, and F1 score have been measured.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1\,Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{5}$$

4.3. Experimental results and discussion

This section presents Five types of experiments, firstly: recall, precision, and F1 score were have been measured for the proposed approach using hybrid features(LDPD + SIFT), then remeasure recall, precision, and F1 score for SIFT, LDPD separately. Secondly: compare the average of detection rate between (LDPD+SIFT) with SIFT and LDPD. Thirdly: we have repeated the experiments on the same objects using other features with LDPD like Haar+LDPD, Raw+LDPD, and Histogram+LDPD to compare their detection rate with proposed approach using SIFT+LDPD features. Fourthly: displaying the effectiveness of sliding window approach in the proposed approach. Finally, we have compared different classifiers with the proposed approach. Table 1 shows how the proposed approach using SIFT+LDPD is outperformed than using solely SIFT or Solely LDPD as displayed in Tables 2 and 3.

Table 1. Recall, precision and F1 score for object detection in the proposed approach

Object Name	Recall %	Precision %	F1 score %
Bowls	84.21	94.12	88.89
Caps	84.52	81.61	83.04
Cereal Boxes	92.86	97.85	95.29
Coffee Mugs	91.58	94.57	93.05
Soda Cans	86.81	89.77	88.27
Bell Pepper	88.89	79.12	83.72
Food Plate	93.41	90.43	91.89

Table 2. Recall, precision and F1 score using solely LDPD features

Object Name	Recall %	Precision %	F1 score %
Bowls	43.75	63.64	51.85
Caps	51.90	66.13	58.16
Cereal Boxes	65.85	75.00	70.13
Coffee Mugs	67.65	58.97	63.01
Soda Cans	57.50	69.70	63.01
Bell Pepper	45.45	60.34	51.85
Food Plate	57.53	60.87	59.15

Table 3. Recall, precision and F1 score using solely SIFT features

Object Name	Recall %	Precision %	F1 score %
Bowls	64.71	78.57	70.97
Caps	59.04	74.24	65.77
Cereal Boxes	79.35	90.12	84.39
Coffee Mugs	85.19	78.41	81.66
Soda Cans	70.79	85.14	77.30
Bell Pepper	60.81	63.38	62.07
Food Plate	70.45	83.78	76.54

On the other hand, we have compared SIFT+LDPD with other features merged with LDPD like (Haar+LDPD, Raw+LDPD, and Histogram+LDPD) as displayed in Figure 9 that better results achieved more than used other hybrid features. Moreover, Table 4 shows the contrast in average of recall, precision, and F1 score between using SIFT+LDPD and using solely SIFT or LDPD. Finally, Table 5 demonstrates the effect of using the sliding window technique to improve results by comparing it with results without using sliding window approach.

Table 4. Comparing between the proposed approach across solely SIFT and solely LDPD

Object Name	Recall %	Precision %	F1 score %
The proposed approach using SIFT + LDPD	88.90	89.64	89.16
SIFT	70.05	79.09	74.10
LDPD	55.66	64.95	59.60

Table 5. Comparing between the proposed approach and without using sliding window approach

Object Name	Recall %	Precision %	F1 score %
SIFT + LDPD using sliding window approach	88.90	89.64	89.61
SIFT + LDPD without using sliding window approach	66.32	60.54	63.30

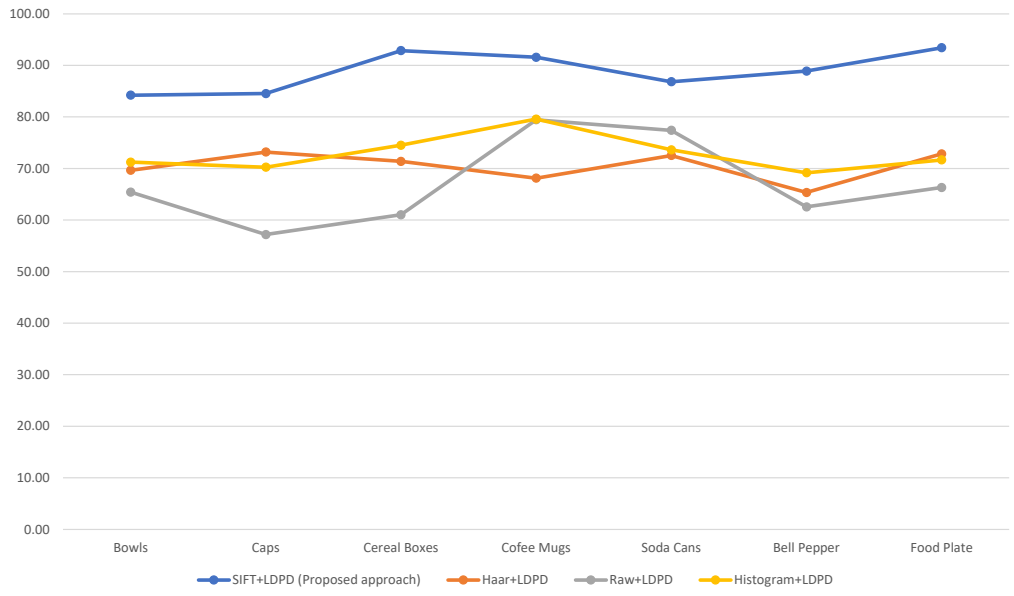

Figure 9. Comparisons between the proposed approach with other features

The results mentioned above proved that the proposed approach using sliding window and hybrid features (SIFT+LDPD) achieve higher results in object detection field more than using other approaches that use other types of features. Moreover, Table 6 is an evidence to prove that the different types of classifiers doesn't really affect on the proposed approach experiment results.

Table 6. Comparing between the average for all objects in terms of different classifiers

Object Name	Recall %	Precision %	F1 score %
Binary SVM + proposed approach	88.90	89.64	89.61
Boosting + proposed approach	86.00	88.10	87.04
Deep Learning + proposed approach	88.75	90.02	89.38

5. CONCLUSION

This paper presents a new approach for object detection in RGBD scenes by merging two types of local features (SIFT and LDPD). In this paper, we have applied our experiments, as well as our proposed approach on the RGB-D object dataset. The experimental results proved that our approach gives us more accurate results in terms of recall, precision, F1 score, and detection rate by using hybrid features, than using one type of feature.

Moreover, the experiments proved that SIFT with LDPD gives better results than using other features with LDPD like Haar, Raw, or Histogram. Also, using crossing sliding window approach improved the results more than using other techniques. Thus, the experiments have displayed that the proposed approach (using hybrid features, crossing sliding windows, binary SVM) is very practical and effective for object detection in RGBD scenes. In the future, authors will improve the proposed approach by using dimensionally reduction technique to compare the results with current proposed approach.

REFERENCES

[1] A. Vahab, M. S. Naik, P. G. Raikar, and S. Prasad, "Applications of object detection system," *International Research Journal of Engineering and Technology (IRJET)*, vol. 6, no. 4, pp. 4186–4192, 2019.

[2] K. Kirkpatrick, "3d sensors provide security, better games," *Commun. ACM*, vol. 61, no. 6, p. 15–17, May 2018, doi: 10.1145/3204449.

[3] D. K. Prasad, "Survey of the problem of object detection in real images," *International Journal of Image Processing (IJIP)*, vol. 6, no. 6, pp. 441-466, 2012.

[4] M. A. Treiber, *An introduction to object recognition: selected algorithms for a wide variety of applications*. UK: Springer-Verlag London. Springer Science & Business Media, 2010, doi: 10.1007/978-1-84996-235-3.

[5] P. Rajput, S. Mittal, and S. Narayan, "Improving accuracy and efficiency of object detection algorithms using multiscale feature aggregation plugins," in *IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR)At: Winterthur, Switzerland*, 2020, pp. 65–76.

[6] S. Krig, "Local feature design concepts, classification, and learning," in *Computer Vision Metrics*, pp. 131–189, 2014, doi: 10.1007/978-1-4302-5930-5_4.

[7] I. I. Ganapathi and S. Prakash, "3d ear recognition using global and local features," *IET Biometrics*, vol. 7, no. 3, pp. 232–241, 2018, doi: 10.1049/iet-bmt.2017.0212.

[8] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Computer Vision and Pattern Recognition*, 2020. [Online]. Available: https://arxiv.org/abs/2006.10511

[9] B. G. Batchelor, *Machine Vision Handbook*. UK: Springer, London, 2012, doi: 10.1007/978-1-84996-169-1.

[10] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *Computer Vision and Pattern Recognition*, 2019. [Online]. Available: http://arxiv.org/abs/1909.00169

[11] K. Lai, L. Bo, and D. Fox, "Unsupervised feature learning for 3d scene labeling," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 3050-3057, doi: 10.1109/ICRA.2014.6907298.

[12] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE international conference on robotics and automation*, 2011, pp. 1817-1824, doi: 10.1109/ICRA.2011.5980382.

[13] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212-3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[14] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3d object recognition in cluttered scenes with local surface features: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2270–2287, 2014, doi: 10.1109/TPAMI.2014.2316828.

[15] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," in *2017 36th Chinese Control Conference (CCC)*, 2017, pp. 11104-11109, doi: 10.23919/ChiCC.2017.8029130.

[16] H. Du, S. Zhao, D. Zhang, and J. Wu, "Novel clustering-based approach for local outlier detection," in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2016, pp. 802-811, doi: 10.1109/INFCOMW.2016.7562187.

[17] T. Zhou, D. Fan, M. Cheng, J. Shen, and L. Shao, "RGB-D salient object detection: A survey," *Comp. Visual Media*, vol. 7, pp. 37–69, 2021, doi: 10.1007/s41095-020-0199-z.

[18] X. Xu, Y. Li, G. Wu, and J. Luo, "Multi-modal deep feature learning for rgb-d object detection," *Pattern Recognition*, vol. 72, pp. 300–313, 2017, doi: 10.1016/j.patcog.2017.07.026.
[19] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 437–451, 2018, doi: 10.1177/0278364917713117.
[20] Y. Zhong, J. Wang, J. Peng, and L. Zhang, "Boosting weakly supervised object detection with progressive knowledge transfer," in *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science,*. vol 12371. Springer, Cham, pp. 615–631, 2020, doi: 10.1007/978-3-030-58574-7_37.
[21] M. Yan, Z. Li, X. Yu, and C. Jin, "An end-to-end deep learning network for 3d object detection from rgb-d data based on hough voting," *IEEE Access*, vol. 8, pp. 138810-138822, 2020, doi: 10.1109/ACCESS.2020.3012695.
[22] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency." Georgia Institute of Technology, 2013. [Online] Available: https://smartech.gatech.edu/handle/1853/51587.
[23] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of international conference on internet multimedia computing and service*, 2014, pp. 23–27, doi: 10.1145/2632856.2632866.
[24] A. F. Otoom, H. Gunes, and M. Piccardi, "Automatic classification of abandoned objects for surveillance of public premises," in *2008 Congress on Image and Signal Processing*, 2008, pp. 542-549, doi: 10.1109/CISP.2008.688.
[25] M. Schwarz and S. Behnke, "Data-efficient deep learning for rgb-d object perception in cluttered bin picking," in *Warehouse Picking Automation Workshop (WPAW), IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2-4.
[26] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.
[27] A. Vit and G. Shani, "Comparing rgb-d sensors for close range outdoor agricultural phenotyping," *Sensors*, vol. 18, no. 12, 2018, doi: 10.3390/s18124413.
[28] R. Wang, W. Wan, Y. Wang, and K. Di, "A new rgb-d slam method with moving object detection for dynamic indoor scenes," *Remote Sensing*, vol. 11, no. 10, 2019, doi: 10.3390/rs11101143.
[29] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing training data for object detection in indoor scenes," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1702.07836
[30] Z. Liao, W. Wang, X. Qi, and X. Zhang, "Rgb-d object slam using quadrics for indoor environments," *Sensors*, vol. 20, no. 18, 2020, doi: 10.3390/s20185150.
[31] J. Chen, B. Lei, Q. Song, H. Ying, D. Z. Chen, and J. Wu, "A hierarchical graph network for 3d object detection on point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 389-398, doi: 10.1109/CVPR42600.2020.00047.
[32] J. Lahoud and B. Ghanem, "2d-driven 3d object detection in rgb-d images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4632-4640, doi: 10.1109/ICCV.2017.495.
[33] Z. Yong, Z. Xiaoxia, and D. Nana, "Research on 3d object detection method based on convolutional attention mechanism," in *Journal of Physics: Conference Series*, 2021, vol. 1848, pp. 012097(1-7), doi: 10.1088/1742-6596/1848/1/012097.
[34] Z. Liu, J. Tang, Q. Xiang, and P. Zhao, "Salient object detection for rgb-d images by generative adversarial network," *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 25403–25425, 2020, doi: 10.1007/s11042-020-09188-8.
[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
[36] S. Awwad, F. Hussein, and M. Piccardi, "Local depth patterns for tracking in depth videos," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1115–1118, doi: 10.1145/2733373.2806295.
[37] S. Awwad and M. Piccardi, "Local depth patterns for fine-grained activity recognition in depth videos," in *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2016, pp. 1-6, doi: 10.1109/IVCNZ.2016.7804453.
[38] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: https://mlab.cb.k.u-tokyo.ac.jp/ moris/lecture/cb-mining/4-svm.pdf