A Computer Vision System for Street Sweeper Robot

Ouiem Bchir, Sultana Almasoud, Lina Alyahya, Renad Aldhalaan, Lina Alsaeed, Nada Aldalbahi
Department of Computer Science, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia

***Abstract*—With the spread of Covid-19, more people wear personal protective equipment such as gloves and masks. However, they are littering them all over streets, parking lots and parks. This impacts the environment and damages especially the marine ecosystem. Thus, this waste should not be discarded in the environment. Moreover, it should not be recycled with other plastic materials. Actually, they have to be separated from regular trash collection. Furthermore, littering gloves and masks yields more workload for street cleaners and presents potential harm for them. In this paper, we design a computer vision system for a street sweeper robot that picks up the masks and gloves and disposes them safely in garbage containers. This system relies on Deep Learning techniques for object recognition. In particular, three Deep Learning models will be investigated. They are: You Only Look Once (YOLO) model, Faster Region based Convolutional Neural Network (Faster R-CNN) and DeepLab v3+. The experiment results showed that YOLO is the most suitable approach to design the proposed system. Thus, the performance of the proposed system is 0.94 as F1 measure, 0.79 as IoU, 0.94 as mAP, and 0.41 s as Time to process one image.**

Keywords*—Covid-19; street sweeper robot; personal protective equipment (PPE); computer vision; deep learning

I. Introduction

In 2019, the Covid-19 pandemic started. It began to spread widely in early 2020, and was classified pandemic by the World Health Organization on March 11, 2020 when the number of infected cases reached 118,319 cases [1]. Since the infection may be spread by a person's sneezing, coughing, spitting and breathing, most countries' governments have imposed wearing face masks in public places and gatherings. In addition, supermarkets impose the use of gloves as well. This led to a significant increase in the use of face masks and gloves. As a consequence, globally, people are using and disposing of approximately 129 billion face masks and 65 billion gloves every single month during Covid-19 pandemic [2]. Unfortunately, many people are throwing these masks and gloves everywhere such as streets, parking lots, gardens and sidewalks. As a result, they will end up in the ocean through sewer systems creating a new form of pollution. In fact, they will shatter into micro plastics and will be contaminated by dangerous chemicals. Moreover, littering gloves and masks lead to a heavier workload for street cleaners. Furthermore, masks and gloves waste are dangerous for the cleaners' health since they are potentially infected. The same problem encountered by street waste workers, is also faced by recycling waste workers. In addition, Personal Protective Equipment (PPE) cannot be sorted with other material in the recycling centers [3]. In fact, they are thin and easily broken and can block and break down the sorting machine. Therefore, PPE waste materials have to be placed in separate sealed bags or safely tight garbage containers. Nowadays, it is common to incorporate specialized robots to support workers. These machines lessen the workload since they are able to perform repetitive and simple tasks efficiently. In particular, a street sweeper robot would alleviate the burden of the waste and recycling workers by picking gloves and masks and storing them in sealed containers. In order to make the robots intelligent and aware of their surrounding environment, integrating sensors with robotics is needed. Specifically, computer vision systems that capture images of the scene surrounding the robot and recognize their content, provide the robot with useful information and an understanding of the scene. In particular, the computer vision system of the street sweeper robot would localize masks and gloves. Typically, suitable visual descriptors should be extracted from the captured images in order to segment the image into several objects. Then, another set of features is extracted from each object in order to recognize it using a classifier. Nevertheless, choosing the suitable feature for the segmentation and the recognition task is not straightforward. In fact, it is one of the main difficulties faced by computer vision systems. Recently, the use of Deep Learning (DL) models alleviated this problem by learning suitable features while training the model. The main goal of this paper is to design and implement a computer vision system for a street sweeper robot that recognizes masks and gloves. This system relies on Deep Learning techniques to recognize objects based on their visual properties. For this purpose, we intend to compare three approaches: You Only Look Once (YOLO) model [4], Faster Region based Convolutional Neural Network (Faster R-CNN) [5], and DeepLab v3+ [6].

II. Background

Object recognition is a field of computer vision which localize and categorize objects in images or video frames. It has been employed in many applications such as tumor recognition in medical images [7], face recognition [8], robot navigation [9], self-driving vehicles [10], etc. Generally, object recognition approaches can be either based on conventional machine learning and image processing techniques or based on Deep Learning approaches. In conventional approaches, a selected set of visual descriptors is extracted from the image for the purpose of segmenting the image into meaningful parts. Then, from the object of interest, another selected feature is extracted and conveyed to a classifier to decide on the class of the object.

Alternatively, object recognition Deep Learning approaches are based on CNN. In fact, they use the conventional layers to 1) automatically learn and extract suitable visual descriptors, and to 2) learn the location of the object. Region Conventional

Neural Networks (R-CNN) models are a well-known family for object recognition. It includes R-CNN [11], Fast R-CNN [12], and Faster R-CNN [5]. Each one of these approaches is an improvement of the previous one. R-CNN is based on a region proposal algorithm called "Selective search". It selects 2000 regions from the image. From each region, visual descriptors are automatically extracted using convolutional layers. Finally, each region is classified using one versus all SVM [13] classifier. In order to enhance the time complexity of the model, Fast R-CNN is proposed. Instead of extracting the visual descriptors from the 2000 regions, visual descriptors are extracted from the whole image first. Then, a Region of Interest (ROI) pooling layer is used to pool the visual descriptors of the region of interest from the final feature map. A SoftMax layer finally classifies this region. An extension to R-CNN [11] and Fast R-CNN [12] is Faster R-CNN [5]. It replaces the "Selective search" algorithm by a Region Proposal Network (RPN). In fact, instead of unnecessarily extracting a fixed number of regions that can be empty or include only a part of the object, Faster R-CNN [5] learns the location of the region to be proposed through the use of a small CNN called RPN. These region-based approaches provide two outputs. These are the bounding boxes coordinate that fits the object of Interest and the class of the object.

Instead of using a region proposal module, Single Shot Detectors (SSDs), use a set of predefined anchor points. From each anchor point, a predefined number of bounding boxes are defined. Then, these models learn if the bounding box contains an object or not, predict the offset of the box so it fits tightly the object, and compute the class probability of each object. Finally, the potential recognized objects are pruned to avoid duplicated recognition. There are various SSDs approaches. They differ in the way of defining the anchors. The most well-known SSD model is YOLO [4].

Another way of semantically understanding the scene is through semantic segmentation. The latter is inextricably related to object recognition. However, it differs in that it does not predict the class and the bounding box of the object, but it learns the pixels that form the object [14]. In fact, semantic segmentation entails assigning a semantic category to each pixel in the input picture.

Recent advances in the field of Deep Learning boosted the semantic segmentation research [15]. In fact, the automatic learning of the features through the convolution layers has improved the performance of semantic segmentation approaches. Nevertheless, CNN cannot be used as it is for semantic segmentation. In fact, max pooling and striding that are suitable for feature reduction, induce low feature resolution [16]. Moreover, since objects may be represented with different scales, standard CNN models need to be trained with different scales of the same object [17]. Furthermore, CNN models discard the location information [18]. Therefore, specific Deep Learning architectures for semantic segmentation have been proposed in the literature. Among these approaches, DeepLab v3+ [6] is a well-known Deep Learning approach for semantic segmentation which has been proven to be effective in many applications [19].

III. Related Works

Due to the Covid-19 pandemic and the need to check if people are wearing the required personal protective equipment (PPE), several masks and gloves recognition systems based on Deep Learning have been reported in the literature [20] [21] [22] [23] [24]. However, no existing work addressed the problem of recognizing masks and gloves thrown in the street. Alternatively, two works based on Deep Learning tackled the problem of recognizing different types of wastes littered in the street [25] [26].

A. Detection of Masks and Gloves Worn by People

The authors in [20] used two Deep Learning models, YOLO (You Only Look Once) [4] and Single-Shot multibox Detector (SSD) MobileNet [27], for the detection and proper wearing of face masks and gloves. First, the model splits the input image into an S × S grid. After that, the grid containing the center of the ground truth bounding box of an object is activated for the detection. Finally, each grid is responsible for predicting the confidence scores of a number of bounding boxes. The MobileNet architecture has been used as a feature extractor in the SSD MobileNet based approach after combining normal convolution and depthwise convolution. The proposed recognition system considers five categories. Namely, it recognizes if the people are wearing masks, not wearing masks, wearing gloves, not wearing gloves, and if they are not properly wearing the masks. The two Deep Learning models were investigated using a dataset containing 8250 photos collected from the internet. The experimental result showed that the proposed system reached an accuracy of 90.6% when using YOLO [4], and an accuracy of 85.5% when using SSD MobileNet [28].

Similarly, the approach in [21] used two Deep Learning models ResNet-50 [29] and YOLOv2 [30]. Nevertheless, they first used ResNet-50 to extract the visual feature. Then, they used YOLOv2 to recognize facial masks. For the assessment of the proposed approach, two medical face masks datasets, Medical Masks dataset (MMD) [31] and Face Mask dataset (FMD) [32] were merged into a single dataset. These datasets have been augmented before being fed to Resnet-50. It achieved a precision of 81%.

The authors in [22] proposed a Deep Learning approach in order to recognize whether or not people are wearing PPE. More specifically, they adopted the YOLOv4 [33] model. The model has backbone, neck and head parts. In the backbone part, CSPDarknet53 is used as a feature extractor model. In the neck section, Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PAN) are employed. Modified PAN has been used for instance segmentation. For the modification of PAN, they used the concatenation operation instead of the addition operation. The SPP is used to perform max pooling over a feature map. The head part was kept the same as it was in YOLOv3. Four categories have been considered which are, wearing a mask, not wearing a mask, wearing a face shield, and wearing gloves. The authors put together a dataset that includes both collected and captured photos. They collected their own dataset to assess the performance of the system. They obtained a precision of 78% and a recall of 80%.

The study in [23] proposed a mask and face recognition system based on YOLOv3 [34]. First, the videos are recorded using digital cameras. After being processed, they are conveyed to YOLOv3. The latter detects faces and masks. The proposed system was trained on a set of 6,000 photos containing surgical masks which are selected from the MAFA dataset [35]. The experimental results showed that the proposed system achieved an accuracy of 84% in recognizing masks and an accuracy of 96% in recognizing faces. The authors in [24] adopted the VGG-16 [36] Deep Learning model to determine whether or not a person is wearing a facemask and checks if the people in a region are observing physical distance. The model is trained using collected data containing 20,000 images. The input image's height and width are set to 224 pixels. Moreover, data augmentation is employed by applying rotation, rescaling, shifting, and zooming operations. It achieved an accuracy of 97%.

B. General Waste Littered in the Street

The authors in [25] proposed a computer vision system for waste littering quantification. The proposed system is based on a Deep Learning model to localize and classify different types of wastes. They employ the OverFeat-GoogLeNet [37] Deep Learning model. It is an adaptation of the OverFeat model [38] which uses GoogLeNet [39] model as backbone deep network model. The authors collected their own dataset using a high-resolution camera placed on the top of a vehicle to take pictures of wastes on the streets and sidewalks. The performance of the system on the 18,676 collected images is 77.35% for the precision and 60% for the recall. Alternatively, the authors in [26] proposed a robot system that is able to pick up garbage from the grass independently. The computer vision part of the robot aims at recognizing general waste using ResNet-34 [29] and SegNet [40] Deep Learning models. The input image is first segmented using SegNet. The latter is a Deep Learning model designed for segmentation. It is based on a decoder-encoder model where the input image is first down-sampled to learn the visual descriptor, then the obtained visual descriptor is up-sampled to recover the input image resolution. After that, the segmented objects are conveyed to ResNet [29] in order to categorize the waste. In fact, the system considers six categories. Specifically, five classes are used for the waste and one class for non-waste. The system is trained on 40k training pictures and tested on 7k testing pictures. Moreover, they collected 750 more pictures representing non-waste for testing. Experiments have shown that the accuracy of littered waste recognition reached up to 95%.

IV. System for Street Sweeper Robot

In order to design a computer vision system that recognizes masks and gloves, we intend to compare the performance of the three approaches: YOLO [4], Faster R-CNN [5], and DeepLab v3+ [6]. First, we need to train the three considered models. In order to train the YOLO [4] model, we feed its input with images representing littered gloves and masks. The labels of these images are also provided to the recognition system to ensure the training. The labels consist of the corresponding categories of the considered objects (gloves and masks), and their surrounding boxes' information, namely, the upper left corner coordinates, the width and height of the box. Similarly, Faster R-CNN [5] is trained in the same way since it uses the same type of labels. Alternatively, DeepLab v3+ [6] employs a different type of labels. In fact, since it is a segmentation approach, the label of each pixel should be provided. More specifically, the captured images with littered masks and gloves are conveyed to the input of DeepLab v3+ [6]. Moreover, their corresponding masks images are provided to the networks. They consist of the same image where the pixels corresponding to each object are manually colored with a different color.

Using YOLO [4], the masks and gloves will be recognized and localized. The obtained results will be assessed to measure the performance of the YOLO [4] based system in terms of the Average Precision per class, Mean Average Precision, IoU, F1 measure and Time to process one image. Similarly, the same procedure will be used for the Faster R-CNN [5] based system. And for the DeepLab v3+ [6] based system, the obtained results will be assessed in terms of the IoU, F1 measure and Time to process one image. In fact, AP and mAP are not defined in the case of semantic segmentation.

When the performances of the three systems are computed, we compare between them in order to conclude on the best system to be considered. We should mention we prioritize the recognition performance over the time one. However, in case the recognition performance is similar or slightly different, we select the faster model. The selected approach among the three considered ones will be adopted as illustrated in Fig. 1.

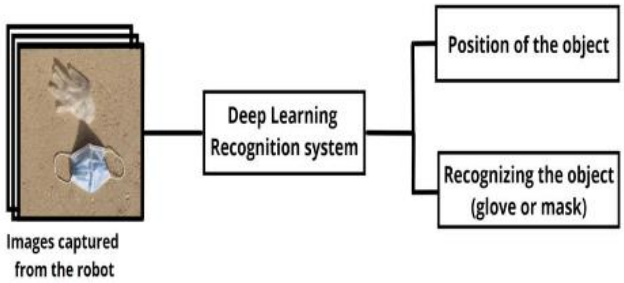


Fig. 1. Proposed System Architecture.

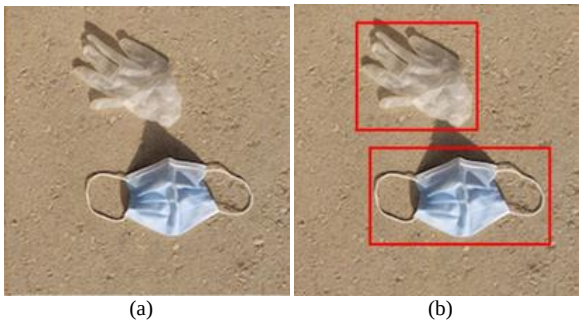
V. Experiments

A dataset of 1500 images containing masks and/or gloves is collected. They are captured using a digital camera and have a size of 224 X 224 pixels. Different backgrounds such as grass, stones, and Asphalt concrete with various angles for shooting and different lighting are considered. Moreover, the masks and gloves in the collected dataset differ in terms of number, color, material, and design. Furthermore, these masks and gloves can be twisted, knotted, or choppy. Two Ground Truth labels are considered. The first one consists of labelling the pixels which belong to the gloves, the masks, and to the background. More specifically, for each image in the dataset, the pixels corresponding to the gloves are colored with green, those corresponding to the masks are colored with blue, and all remaining pixels are colored in black. The coloration is done manually. This first type of Ground Truth will be used with DeepLab v3+ [6] which requires pixel wise labelling. Fig. 2 shows a sample image and the corresponding pixel wise labelling as required by DeepLab v3+ [6].

(a) (b)

Fig. 2. Sample Pixel Wise Labelled Image. (a) The Sample Image, (b) The Corresponding Ground Truth Required by DeepLab v3+.

Alternatively, for YOLO [4] and Faster R-CNN [5] a different type of Ground Truth labelling is required. In fact, these two approaches, require the bounding boxes coordinates of each object of interest and its corresponding class (mask or glove) with respect to each considered image. The coordinates of the bounding box consist of the upper left coordinates, the width and the height of the rectangle surrounding tightly the object. Fig. 3 depicts sample images and the corresponding bounding boxes of the object of interest.



(a) (b)

Fig. 3. Sample Labelled Image with Bounding Boxes. (a) The Sample Image, (b) The Corresponding Bounding Boxes Required by YOLO and Faster R-CNN.

A. *Experiment 1*

This experiment aims at assessing the performance of YOLOv4 [33] to recognize masks and gloves. For this purpose, we want to figure out the best hyper-parameter configuration for YOLOv4 [33]. As such, YOLOv4 model was trained using different values of the learning rate, the momentum and the number of batches. The learning rate is the most crucial hyper-parameter. In fact, a too small value may result in a long training process, whereas a too large value may result in overshooting the global minimum. Table I shows the five considered configurations. In order to determine the best model, the performance results on the validation set of each configuration are reported. They are the IoU, F1 measure, AP, mAP, and Time to process one image.

Fig. 4 shows the performance measures of YOLOv4 [33] model. As shown in Fig. 4, the best performance is obtained when using configuration 3, and the worst performance is obtained when using configuration 4. In fact, the learning rate in configuration 3 is set to 0.001 while the learning rate in configuration 4 is set to 0.1. Thus, the learning rate of 0.1 yielded the overshoot of the optimal model. Moreover, when the number of batches is large, the performance is better. This due to the fact that the prediction error used to update the weights is computed using a larger number of images at each batch. The obtained results are 0.95 as F1 measure, 0.8072 as IoU, and 0.963632 as mAP using the considered configuration. We should note that configuration 4 performed better in terms of the time to process one image. However, we prioritize the recognition performance over the time one. Moreover, the difference is not significant, we prioritize the recognition performance. Using configuration 3, the results of the validation, and testing are reported in Table II. As shown, there is no significant performance drop when using the test set. Therefore, we can assume that the learned model is not over-fitted.

In order to illustrate the result obtained by the YOLOv4 [33] model, three sample results are depicted in Fig. 5. In Fig. 5(a), Fig. 5(c), and Fig. 5(e), the images fed to the YOLO model are displayed. Moreover, in Fig. 5(b), Fig. 5(d), and Fig. 5(f), the obtained results are shown. More specifically, the bounding box surrounding the object of interest along with the confidence score are displayed. As shown, even if the gloves or masks are folded, or overlap with another object, YOLOv4 [33] model is able to recognize them with a high confidence score.

TABLE I. YOLOV4 [33] HYPER-PARAMETER CONFIGURATIONS

	Learning rate	**Momentum**	**Number of batches**
Configuration 1	0.001	0.949	64
Configuration 2	0.01	0.95	128
Configuration 3	0.001	0.949	256
Configuration 4	0.1	0.99	128
Configuration 5	0.001	0.949	8

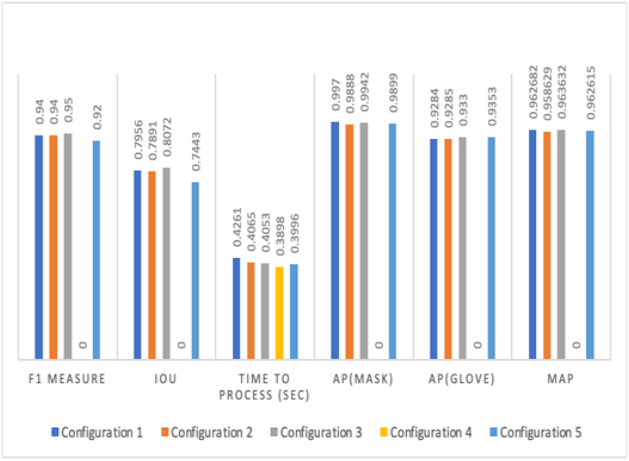


Fig. 4. YOLOv4 [33] Performance Results.

TABLE II. PERFORMANCE OF YOLOV4 [33] MODEL ON THE VALIDATION AND TEST DATASETS

	F1 Measure	**IoU**	**Time to process (in sec)**	**AP (Mask)**	**AP (Glove)**	**mAP**
Validation Set	0.95	0.81	0.41	0.99	0.93	0.96
Test Set	0.94	0.79	0.41	0.98	0.90	0.94

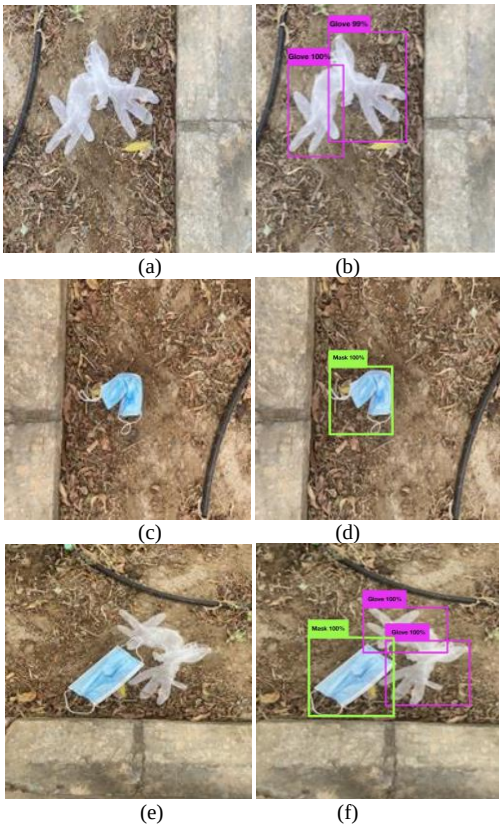


Fig. 5. Three Sample Results Illustrating the Result Obtained when using YOLOv4 [33] Model. (a) Sample Image 1, (b) The Output of Sample Image 1, (c) Sample Image 2, (d) The Output of Sample Image 2, (e) Sample Image 3, (f) The Output of Sample Image3.

B. Experiment 2

In this experiment, we evaluate the performance of Faster R-CNN [5]. In this regard, the hyper-parameters are tuned using the validation set. Table III shows the five considered configurations.

Fig. 6 shows the performance measures of Faster R-CNN [5] on the validation set with respect to each considered configuration. As shown in Fig. 6, the best result is obtained using configuration 3. Specifically, the hyper-parameters are set to 0.0001 for the learning rate, 0.96 for the momentum and 690 for the number of batches. We should mention that configuration 5 gave slightly better AP with respect to the Mask class, but not with respect to the Glove class. However, in terms of mAP configuration 3 is better. The corresponding performance results are 0.5665 as F1 measure, 0.7337 as IoU, and 0.4350 as mAP. Table IV reports Faster R-CNN performance on both the validation and test sets. As shown, the performance of the test set is not significantly worse than the performance of the validation set, and thus the overfitting assumption is discarded.

Fig. 7 displays three sample results of the Faster R-CNN model. The images conveyed to Faster R-CNN model are displayed in the Fig. 7(a), Fig. 7(c), and Fig. 7(e), respectively. The corresponding output results are shown in Fig. 7(b), Fig. 7(d), and Fig. 7(f), respectively. As shown in Fig. 7, the bounding box surrounding the object of interest along with the confidence score are depicted. We can notice that Faster R-CNN is able to recognize the withdrawn gloves and masks. Nevertheless, for some cases such the illustrative example displayed in Fig. 7(d), the confidence score is not high. This can be due to the fact the mask is folded.

C. Experiment 3

In order to detect masks and gloves using semantic segmentation, we tuned the hyper-parameters for DeepLab v3+ [6]. In this regard, we trained DeepLab v3+ [6] using ResNet-50 [29] as backbone. Then, using the validation set, we evaluated the model with respect to five considered configurations. In particular, the learning rate, momentum and number of batches were tuned. These five configurations are reported in Table V.

TABLE III. FASTER R-CNN [5] HYPER-PARAMETER CONFIGURATIONS

	Learning rate	Momentum	Number of batches
Configuration 1	0.001	0.96	690
Configuration 2	0.001	0.94	690
Configuration 3	0.0001	0.96	690
Configuration 4	0.00001	0.96	690
Configuration 5	0.0001	0.94	690

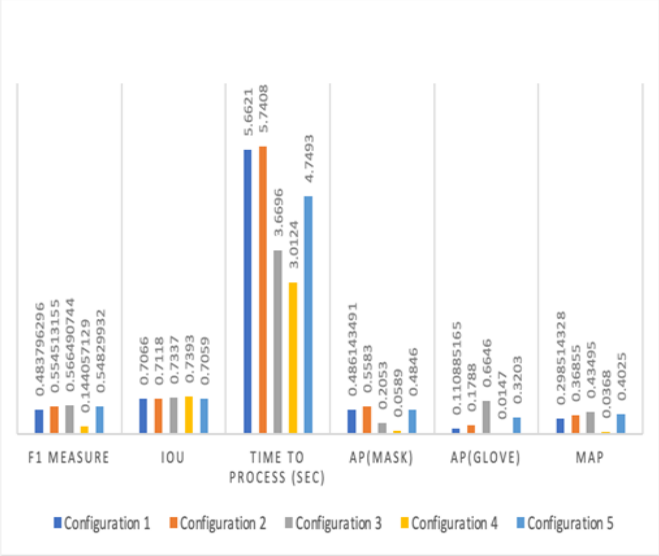


Fig. 6. Faster R-CNN [5] Performance Results.

TABLE IV. PERFORMANCE OF FASTER R-CNN [5] MODEL ON THE VALIDATION AND TEST DATASETS

	F1 Measure	IoU	Time to process (in sec)	AP (Mask)	AP (Glove)	mAP
Validation Set	0.57	0.73	3.67	0.21	0.66	0.43
Test Set	0.50	0.74	3.69	0.21	0.45	0.33

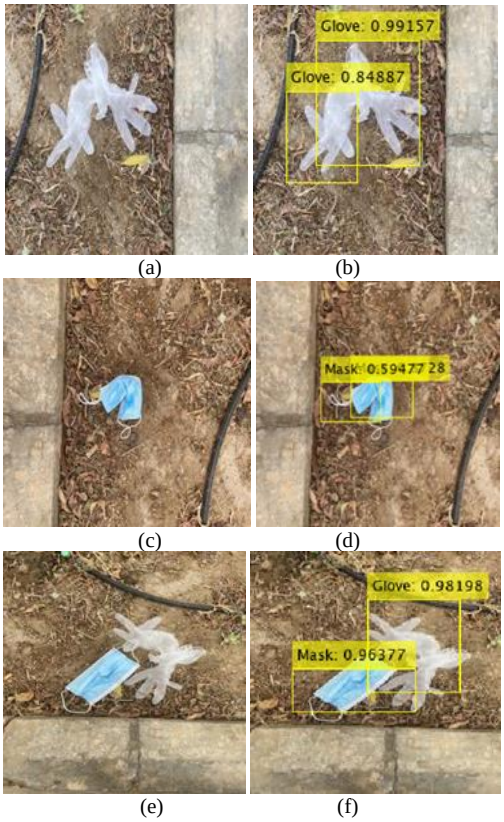


Fig. 7. Three Sample Results Illustrating the Result Obtained using Faster R-CNN Model. (a) Sample Image 1, (b) The Output of Sample Image 1, (c) Sample Image 2, (d) The Output of Sample Image 2, (e) Sample Image 3, (f) The Output of Sample Image 3.

TABLE V. DEEPLAB V3+ [6] HYPER-PARAMETER CONFIGURATIONS

	Learning rate	Momentum	Number of batches
Configuration 1	0.001	0.95	690
Configuration 2	0.01	0.97	690
Configuration 3	0.1	0.99	345
Configuration 4	0.001	0.96	171
Configuration 5	0.001	0.99	690

Fig. 8 displays the performance measures of the system when using DeepLab v3+ [6] semantic segmentation approach with ResNet-50 [29] as backbone. We should notice that contrary to the previous two experiments, only F1 measure, IoU and Time to process are considered. In fact, Average precision performance measure is not defined for segmentation approaches. As shown in Fig. 8, using configuration 1, a learning rate of 0.001, momentum of 0.95 and number of batches of 690, yielded the best performance result with an IoU of 0.9762, and F1 measure of 0.98. In fact, configuration 1 is characterized by a small learning rate avoiding missing the global minimum of the error rate, a large batch size enhancing the error computation, and a relatively smaller momentum (percentage of previous iteration gradients to be considered). Alternatively, a large learning rate of 0.1 (configuration 3) gave the worst result. This can be explained by an under-fitting situation where the model fails to find the global minimum and converges early since the learning step is too large.

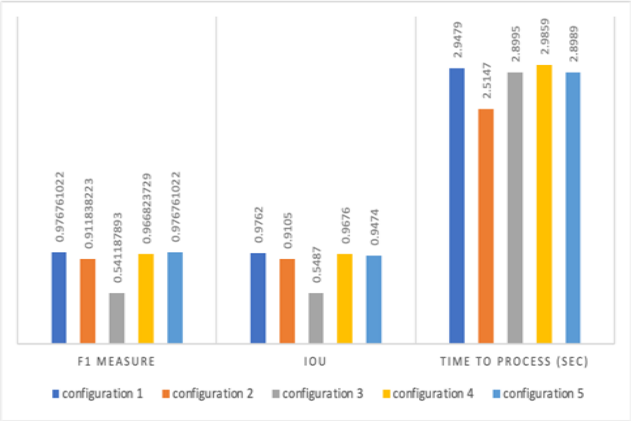


Fig. 8. DeepLab v3+ [6] Performance Results.

Using configuration 1, we evaluated the performance of DeepLab v3+ on the test images. Table VI depicts the performance results of both the validation and the test sets. As reported, there is no drop in the performance when using the test set compared with the performance of the validation set. Therefore, we can conclude that the learned model is not over-fitted.

TABLE VI. PERFORMANCE OF DEEPLAB V3+ [6] MODEL ON THE VALIDATION AND TEST DATASETS

	F1 Measure	IoU	Time to process (in sec)
Validation Set	0.98	0.98	2.95
Test Set	0.99	0.98	3.01

Fig. 9 displays three sample results of DeepLab v3+ model. The input images are shown in Fig. 9(a), Fig. 9(c), and Fig. 9(e), while the corresponding segmented images are depicted in Fig. 9(b), Fig. 9(d), and Fig. 9(f), respectively. In the segmented images, the pixels recognized as masks by DeepLab v3+ are colored in blue, those recognized as gloves with green, and all remaining pixels belonging to the background in black. We can notice that DeepLab v3+ is able to recognize the pixels belonging to withdrawn gloves and masks in different backgrounds, and for various colors of the gloves and the masks.

D. Discussion

Fig. 10 compares the performances of YOLOv4 [33] and Faster R-CNN [5] in terms of AP and mAP. We should notice that AP and mAP aren't defined for semantic segmentation approaches such as DeepLabv3+ [6]. As shown Fig. 10, YOLOv4 outperforms Faster R-CNN in recognizing both gloves and masks. This is also confirmed by the mAP. In fact, it is higher for YOLOv4 than Faster R-CNN. This can be explained by the fact that YOLO uses a single end-to-end network while Faster R-CNN uses two networks. Therefore, this can reduce the error. Moreover, Faster R-CNN is a region based approach where the classification is performed on the selected region only; whereas YOLO employs the whole image to predict the location and class of the object of interest. Thus, YOLO accesses more contextual information and predicts less false positives of the background. Furthermore, since YOLO has one object rule that makes it predict a single object per cell, it encourages the spatial diversity of the detected objects.

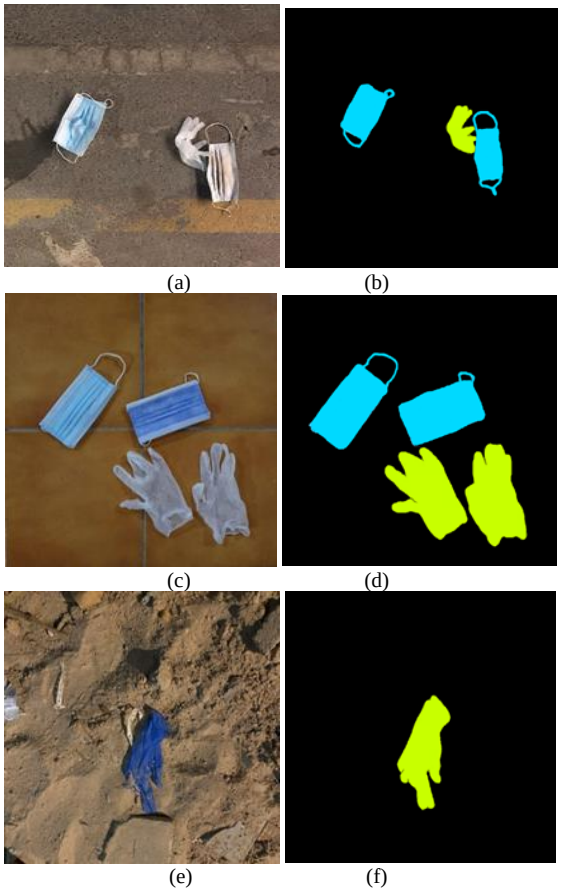


Fig. 9. Three Sample Results Illustrating the Result Obtained when using DeepLab v3+ Model. (a) Sample Image 1, (b) The Output of Sample Image 1, (c) Sample Image 2, (d) The Output of Sample Image 2, (e) Sample Image 3, (f) The Output of Sample Image 3.

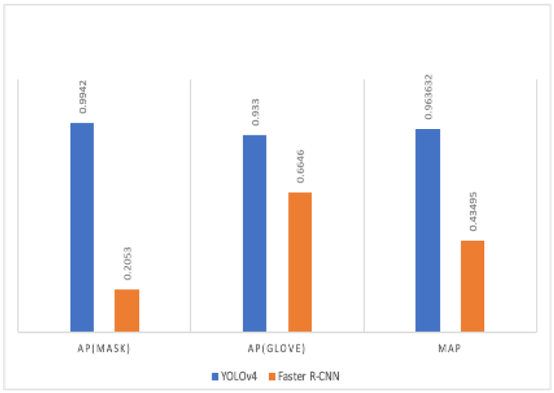


Fig. 10. Performance Comparison of YOLOv4[33], and Faster R-CNN [5] in Terms of AP, and mAP.

Fig. 11 compares the performance of YOLOv4 [33], Faster R-CNN [5], and DeepLab v3+ [6] in terms of F1 measure, IoU, and Time to process one image. As depicted, DeepLab v3+ is better in localizing masks and gloves with an IoU equal to 0.98, compared to an IoU equal to 0.81 for YOLOv4, and 0.73 for Faster R-CNN. This is an expected result since semantic segmentation is a more powerful approach for localizing the object of interest since it works at the pixel level, and not the bounding box like YOLOv4 and Faster R-CNN. Moreover, DeepLab v3+ is slightly outperforming YOLOv4 and Faster R-CNN according to the F1 measure which combines both the localization and the classification performances of the object of interest. Nevertheless, in terms of processing time, YOLOv4 highly outperforms the other approaches with a time to process one image equal 0.41 s against 3.7 s for Faster R-CNN and 2.95 s for DeepLab v3+. Since the F1 measure difference between YOLOv4 and DeepLab v3+ is not significant while the difference in terms of processing time is large in favor of DeepLab v3+, we choose the YOLOv4 model to design the proposed approach.

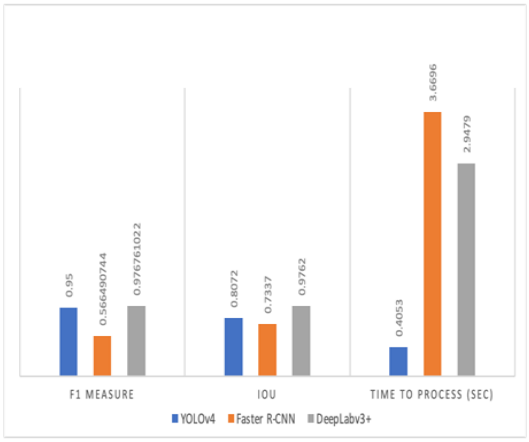


Fig. 11. Performance Comparison of YOLOv4, Faster R-CNN, and DeepLab v3+ in Terms of F1 Measure, IoU, and Time to Process One Image.

In attempt to further enhance the performance of the selected model (YOLOv4), we employed data augmentation. In other words, additional images are considered to train the model. These images are obtained by modifying existing images using rotation, cropping, blurring and adding brightness. The augmented dataset consists of 2073 images, Fig. 12 shows sample images from the augmented dataset. Table VII depicts the performance of YOLOv4 when using data augmentation and without using it. We noticed that using the augmented dataset shows a slight improvement to the results in terms of F1 measure, IoU, AP, mAP.

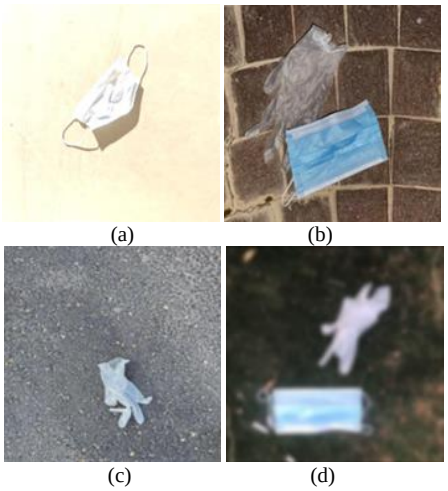


Fig. 12. Sample Images Obtained after the Data Augmentation. (a) Brightened Image, (b) Rotated Image, (c) Cropped Image, and (d) Blurred Image.

TABLE VII. PERFORMANCE COMPARISON OF YOLOV4 [33] WHEN USING DATA AUGMENTATION AND WITHOUT USING IT

	F1 Measure	IoU	Time to process (in sec)	AP (Mask)	AP (Glove)	mAP
Test results with augmented data	0.95	0.81	0.49	0.99	0.94	0.97
Test results without augmented data	0.94	0.79	0.41	0.982	0.901	0.94

E. Conclusion and Future Works

In this paper, we propose to design and implement a computer vision system of a street sweeper robot that recognizes masks and gloves for the purpose of picking them and disposing them in securely tight garbage bags. The proposed system is based on a Deep Learning object recognition approach.

After investigating the related works and studying the related background, we proposed an effective approach to automatically recognize facial masks and gloves which have been littered in the environment. For these purposes, three Deep Learning approaches are compared. These are YOLO, Faster R-CNN and DeepLab v3+. YOLOv4 is selected as the most suitable model for detecting littered gloves and masks. As future works, we propose to investigate emerging deep learning recognition and semantic segmentation approaches. In fact, pattern recognition field is an active field of research with continuous advancement.

REFERENCES

[1] "Coronavirus Disease 2019 (COVID-19) Situation Report -51 SITUATION INNUMBERS Total and New Cases in Last 24 Hours, Mar. 2020." Who.int. https://www.who.int/docs/default-source/coronaviruse/ situationreports/20200311- sitrep51covid-19.pdf?sfvrsn=1ba62e57_10 (accessed Sep. 5, 2021).

[2] S. Hirsh. "Every Month, 200 Billion Face Masks and Gloves Are Going Into the Environment." GreenMatters.com. https://www.greenmatters. com/p/face-masks-gloves-littercoronavirus (accessed Sep. 5, 2021).

[3] S. Waldek. "How to Properly Dispose of PPE." HouseBeautiful.com.https://www.housebeautiful.com/lifestyle/a335767 81/how-to-dispose- ppemasksglovesrecyclable/ (accessed Sep. 12, 2021).

[4] J. Redmon. "YOLO: Real-Time Object Detection." Pjreddie.com.https: //pjreddie.com/darknet/yolov1/ (accessed Sep. 12, 2021).

[5] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time ObjectDetection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, Jun. 2017, doi:10.1109/TPAMI.2016.2577031.

[6] L. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," presented at arXiv, 2021.

[7] E. Mohd. Azhari, M. Mudzakkir Mohd. Hatta, Z. Zaw Htike and S. Lei Win, "Tumor Detection in Medical Imaging: a Survey, " International Journal of Advanced Information Technology (IJAIT), vol. 4, no. 1, pp. 21-30, Feb. 2014, doi: 10.5121/ijait.2014.4103.

[8] J. Weng and W. Hwang, "Toward automation of learning: the state self-organization problem for a face recognizer," in Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, Apr. 14-16, 1998, pp. 384-389, doi: 10.1109/AFGR.1998.670979.

[9] W. Gueaieb and M. S. Miah, "An Intelligent Mobile Robot Navigation Technique Using RFID Technology," in IEEE Transactions on Instrumentation and Measurement, vol. 57, no.9, pp. 1908-1917, Sep. 2008, doi: 10.1109/TIM.2008.919902.

[10] A. Uçar, Y. Demir and C. Güzeliş, "Object recognition and detection with deep learning for autonomous driving applications," in SIMULATION, vol. 93, no. 9, pp. 759-769, Jun. 2017.doi: 10.1177/0037549717709932.

[11] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, Jun. 23-28, 2014, pp. 580- 587, doi: 10.1109/CVPR.2014.81.

[12] R. Girshick, "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec. 7-13, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[13] Yi Liu and Y. F. Zheng, "One-against-all multi-class SVM classification using reliability measures," Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Montreal, Que, Jul. 31-4 Aug., 2005, vol. 2, pp. 849-854, doi: 10.1109/IJCNN.2005.1555963.

[14] J. Shi and L. Zhao, "A Review of Lane Detection Based on Semantic Segmentation", in International Journal of Advanced Network, Monitoring and Controls, vol. 6, no. 3, pp. 1-8, Oct. 2021. [Online]. Available: http://www.ijanmc.org/Uploads/20213/2021-03- 01.pdf.

[15] T. Lei, Z. Jiao and A. Nandi. "Recent Advances in Image and Video Semantic Segmentation Using Deep Learning." Frontiers.org. https://www.frontiersin.org/researchtopics/22286/recent- advances-in-image- andvideosemantic-segmentation-using-deeplearning (accessed Oct. 30, 2021).

[16] J. Brownlee. "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks." MachineLearningMastery.com. https://machine learningmastery.com/poolinglayers-for-convolutional-neural-networks/ (accessed Oct. 30, 2021).

[17] N. van Noord and E. Postma, "Learning scale-variant and scale-invariant features for deep image classification," in Pattern Recognition, Zhongshan, China, Jan. 2017, pp. 583- 592.

[18] M. Ghafoorian et al, "Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities," in Scientific Reports, Jul. 2017, pp. 1– 12.

[19] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in IEEE Transactions on Pattern Analysis and Machine Intelligence, United States, June. 2016, pp. 99-113.

[20] S. Khosravipour, E. Taghvaei, and N. Charkari, "COVID-19 personal protective equipment detection using real-time deep learning methods," in arXiv ,Mar. 2021, pp. 11– 20.

[21] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against covid19: A novel deep learning model based on Yolo-V2 with resnet-50 for medical face mask detection," in Sustainable Cities and Society, Nov. 2020, pp. 1-8.

[22] A. Protik, A. H. Rafi and S. Siddique, "Real-time Personal Protective Equipment (PPE) Detection Using YOLOv4 and TensorFlow," in 2021 IEEE Region 10 Symposium (TENSYMP), 2021, pp. 1-6, doi: 10.1109/TENSYMP52854.2021.9550808.

[23] R. Avanzato, F. Beritelli, M. Russo, S. Russo and M. Vaccaro, "YOLOv3-based mask and face recognition algorithm for individual protection applications," in Search.bvsalud.org, 2021, pp. 41-45.

[24] S. V. Militante and N. V. Dionisio, "Deep Learning Implementation of Facemask and Physical Distancing Detection with Alarm Systems," in 2020 Third International Conferencon Vocational Education and Electrical Engineering (ICVEE), 2020, pp. 1-5, doi: 10.1109/ICVEE50212.2020.9243183.

[25] M. Saeed, A. Kaenel, A. Droux, F. TiEche, N. Ouerhani, H. Ekenel and J. Thiran, "A Computer Vision System to Localize and Classify Wastes on the Streets," in arXiv , 2017, pp 195-204.

[26] J. Bai, S. Lian, Z. Liu, K. Wang and D. Liu, "Deep Learning Based Robot for Automatically Picking Up Garbage on the Grass," in IEEE Transactions on Consumer Electronics, vol. 64, no. 3, pp. 382-389, Aug. 2018, doi: 10.1109/TCE.2018.2859629.

[27] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural

networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[28] T. Ghosh, L. Li, and J. Chakareski, "Effective Deep Learning for Semantic Segmentation Based Bleeding Zone Detection in Capsule Endoscopy Images," 2018 25th IEEE International Conference on Image Processing (ICIP), no. September 2019, pp. 3034– 3038, 2018.

[29] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas,NV, USA, Jun. 27-30, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[30] R.Li and J.Yang, "Improved YOLOv2 Object Detection Model,"2018 6th International Conference on Multimedia Computing and Systems (ICMCS), 2018, pp. 1-6, doi: 10.1109/ICMCS.2018.8525895.

[31] Medical mask dataset, Humans in the Loop, 2021. [Online]. Available: https://humansintheloop.org/resources/datasets/medical-mask-dataset/.

[32] Face mask detection, Kaggle, May. 2020. [Online]. Available: https://www.kaggle.com/andrewmvd/face-mask-detection

[33] A. Bochkovskiy, C. Wang and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," in arXiv.org, Apr. 2020, pp. 1-17.

[34] H. Gong, H. Li, K. Xu and Y. Zhang, "Object Detection Based on Improved YOLOv3- tiny," in 2019 Chinese Automation Congress (CAC), 2019, pp. 3240-3245, doi: 10.1109/CAC48633.2019.8996750.

[35] mafa-dataset, Kaggle, 2021.[Online]. Available: https://www.kaggle.com/revanthrex/mafadataset.

[36] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for LargeScale Image Recognition," presented at the 3rd International Conference on Learning Representations, San Diego. USA , Sep. 2014.

[37] R. Stewart, M. Andriluka and A. Y. Ng, "End-to-End People Detection in Crowded Scenes," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2325-2333, doi: 10.1109/CVPR.2016.255.

[38] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," presented at arXiv, 2021.

[39] C. Szegedy et al, "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[40] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional EncoderDecoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.