Garbot - Semantic Segmentation for Material Recycling and 3D Reconstruction Utilizing Robotics

Siva Ariram, Tuulia Pennanen, Antti Tikanmäki, and Juha Röning
Biomimetics and Intelligent Systems Group,
Faculty of Information Technology and Electrical Engineering,
University of Oulu, Finland
siva.ariram@oulu.fi

***Abstract*—Semantic segmentation directly from the images of landfills can be utilized in the earth movers to segregate the garbage autonomously. Generally, Various segregation methods are available for garbage segregation such as IOT based waste segregation, Conveyor belt segregation in which none of them are directly from landfills. Semantic segmentation is one of the important tasks that maps the path towards the complete scene understanding. The aim of this paper is to present a smart segregation method for garbage by using semantic segmentation with DeepLab V3+ Model using the framework(Backbone model) of Xception-65 with the mean accuracy of 75.01%. This paper features the segmentation with the GarbotV1 dataset which has major classifications such as Plastic, Cart-board, Wood, Metal, Sponge. The paper also contributes a method for reconstructing the segmented images to build a 3D map and this exploits the use of earth moving vehicles to navigate autonomously by localizing the segmented objects.**

***Index Terms*—Semantic segmentation, 3D reconstruction, Landfills, garbage segregation**

I. Introduction

Waste segregation is one of the most salient projects for helping the environment. The segregation of waste still doesn't have a proper solution because the segregation method is not so widely present. We dump a massive 2.01 billion tons of waste every year. There will be a drastic increase of waste by 2050, the world is expected to generate 3.40 billion tons of waste annually [4].In most of the Landfill centres, earth movers do the first stage of separation manually which are time consuming and expensive. The development of on-site equipment in waste management industry to integrate robotics and AI technology to make them autonomous and more efficient.

Therefore, the aim of the paper is to present a segregation method of garbage by segmenting the wastes directly from the landfills through camera using semantic segmentation which can be incorporated into earth moving vehicles to make it autonomous. Image segmentation is a machine vision task where we do label a specific object or region within the frame or image based on how it has to be shown. Here, we have developed a DeepLab V3+ network with the backbone model as Xception-65 to predict different classifications of garbage such as Plastic, Cart-board, Wood, Metal, Sponge. To develop these models which rely on huge amount of training data to achieve their complete potential, the semantic segmentation which has dense nature prediction necessitates a expensive data annotation process. The quality of annotation plays a key role for training better models. The outline of our proposal are summarized below:

- We developed DeepLab V3+ semantic segmentation model using the framework(Backbone model) of Xception-65 model with the mean accuracy of 75.01%.
- We created a semantic segmentation dataset containing 5 major classifications of garbage, especially for segmenting garbage directly from landfills.
- We proposed a 3D reconstruction of the segmented images to build 3D model for autonomous earth moving vehicles.

Most of the successful landfill segregation method requires some level of manual segregation which is tedious task. In advances, RGB Camera and an air stream is used to segregate the waste transported separately on the conveyor belt system [5]. Unfortunately, a huge part of the waste is still collected in the form of Municipal solid waste, this is the reason why many countries seeking for the most effective segregation from the landfills. In search of Municipal waste data in Finland, we found 115,159 tonnes of municipal(Oulu) waste, received during the year 2020 of which 20.93 per cent was recycled as material (2019: 21.69%), 79.01 per cent as energy and other uses (2019: 77.87 %) and 0.06 per cent ended up in final waste disposal (2019: 0.45%) [1].

In IOT based waste segregation [6], Infrared obstacle line sensor is fixed on the outer edge of the trash bin and CNN to classify the garbage into various categories when it goes inside the trash bin. The system is more considering towards collecting the data of garbage which can benefit for its segregation later. Deep Learning techniques for waste segregation [12], The models use SVM with scale invariant feature transform features and a CNN. Convolutional neural network were used to detect the individual trash. In this method, data augmentation techniques were executed on each image because of the size of each class is small.

To understand how semantic segmentation is handled by modern deep learning architectures, it is important to know the steps in progression from coarse to fine inference, which helps in our project to segment the targets directly from landfill images. The prediction of targets in an image or by providing a prior list if there are many of the same targets. After prediction,

Localization or segmentation is the future step towards fine-grained inference, by providing the additional information regarding spatial location of segmented classes. Now it is obvious that semantic segmentation can achieve fine-grained inference to make dense predictions labels for each pixel.

II. Technical Approach

The immense success of deep learning techniques in different high-level machine vision tasks, Semantic segmentation is one of the tasks that maps the roadway towards the complete scene understanding by doing pixel-wise labeling. In this approach, we propose to utilize the video frames of landfills to efficiently segmenting garbage as shown in Fig. 1.

The proposed method consists of three stages: Training the network using backbone model, Testing the model, 3D Reconstruction. In the first stage, the required dataset for the Xception-65 [2] backbone model was annotated and processed for training the DCNN. In the second stage, the DeepLab V3+ model, a semantic segmentation model constructed using Xception-65 [2], was designed to achieve the segmentation of the garbage targets. In the third stage, the simple 3D reconstruction of segmented images been developed for better understanding of 3D coordinates to the output interfaces.

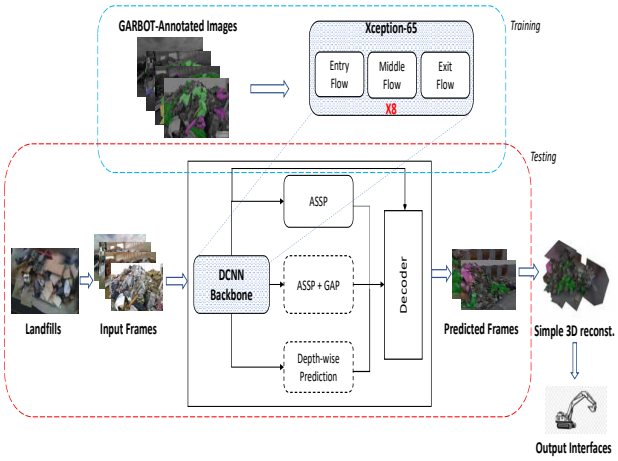

Fig. 1. Architecture. An Xception-65 [2] model is employed in this DCNN for semantic segmentation of garbage. The approach also produces simple 3D Reconstruction that are segmented

A. Data Annotation

Dataset collection is one of the most challenging part and also it plays an important role for training better models in any machine learning systems. This has high priority when dealing with deep networks. For that reason, collection of sufficient data for constructing an appropriate dataset, which has to be fit enough to train models.This dataset pre-processing task, although the simplicity of its formulation in comparison with sophisticated neural network architecture definitions, is one of the hardest problems to solve in this context [7].

Throughout the future, two-dimensional images are going to be utilized for semantic segmentation. On that basis, 2D

Fig. 2. GarbotV1 data samples. Row 1: Original frames. Row 2: Segmented frames. Row 3: Segmented Raw frames

datasets are considered for this garbage segregation. We gathered 9 continuous video frame sequences sparsely annotated at regular intervals. All the images were manually annotated using Labelme: Image polygonal annotation with python [18]. For example, annotating all pixels in a 1024 x 2048 garbage images took average of 15 minutes and some of the sample annotated images are shown in Fig. 2. There are 5 classes categorized into Plastic, Cart-board, Wood, Metal, Sponge and background is also considered when the pixel is not belongs to any of these classes. The GarbotV1 dataset is divided into two subsets as 1498 images for training and 291 images for validation.

B. Semantic Segmentation

For semantic segmentation we employ DeepLabV3+ as it utilizes prominent architectural properties, making it perfectly suitable for garbage segmentation from landfills. A backbone network Xception-65 [2] processes an input image and its output is subsequently segmented by extracting depth-wise pixel labeling. The processing module is either Dense prediction cell [8] or Atrous spatial Pyramid Pooling with or without Global average pooling(GAP) [9]. DeepLabV3+ applies Atrous spatial pyramid pooling, to extract features at different scales, several semantic segmentation architectures performs Spatial pyramid pooling. Atrous convolution [10] is a convolution which integrates spacing between kernel parameters and also it allows us to increase the field of view at any DCNN Layer. To avoid loss of resolution, we assigned its stride to 3 and replaced all subsequent layers with atrous convolutional layers with rate of r=6. This approach converts frame classification networks into dense feature extractors without any requirement of additional parameters. There are two efficient ways to perform atrous convolution, we decided to go with second method, which is to sub-sample the input feature map by a factor equal to the atrous convolution rate r(6, 12, 18, 24), in our case r_h x r_w(6 x 6) possible shifts. The other reason we chose the second method is, its capability

TABLE I
SEMANTIC SEGMENTATION PERFORMANCE(MIOU) BY VARYING BATCH SIZE

Batch Size	mIoU
4	69.42
8	72.64
12	75.01

of applying standard convolution to these intermediate feature maps and tracing them back to original image resolution.

We must make sure that our model is robust to the different size of objects when working with CNN. Atrous Spatial pyramid pooling networks resolved this issue by encoding multi-scale contextual information. The intention towards the multi-scale training is to simulate the varying input sizes while still griping the existing fixed-scale implementations.This approach is well implemented using TensorFlow Framework [11].

C. Backbone Network and Training Protocol

We use the DeepLab v3+ architecture with a network backbone(Xception-65) as a reference and resolved the segregation of it. The observation says the robustness of semantic segmentation models of DeepLab V3+ increases accuracy often with model performance. The Xception based models give significant performance when compared to ResNets and other models, also the ResNet-based backbones are vulnerable when applied for a large-scale dataset [3]. When transferring the results of semantic segmentation, Xception-based models have conclusive structure bias than others(e.g., ResNets). The training protocol of the model as follow:

Learning rate: We employ a polynomial learning rate policy [13],

$$(1 - i/max_i)^p \qquad ,p = 0.9 \qquad (1)$$

Here, i denotes number of iterations, max_i is the total number of iterations and p is the power which has to be increased for better accuracy and faster learning.

Train Crop Size: We fixed crop size to be 513 x 513 for both training and testing on GarbotV1 dataset, large crop size is required for atrous convolution to be effective.

Batch Normalization: Our altered model on top of Xception-65 is included with batch normalization process [14]. It is important to train DeepLab V3+ with fine-tuning batch normalization, We were varying the batch size=(4,8,12) to see the better performance. The batch sizes are entirely with respect to GPU capability.

We observed the quick convergence when using pre-trained weights to train Xception-65 model. The network was trained for comparable amount of time, which resulted in variable number of epochs(Fig. 3).

The garbage segmentation directly from landfills results on Xception-65 is reported in Table I has already tested experiments on the GarbotV1 dataset by varying its batch size for DeepLabV3+, attaining 69.42% with batch size of 4.

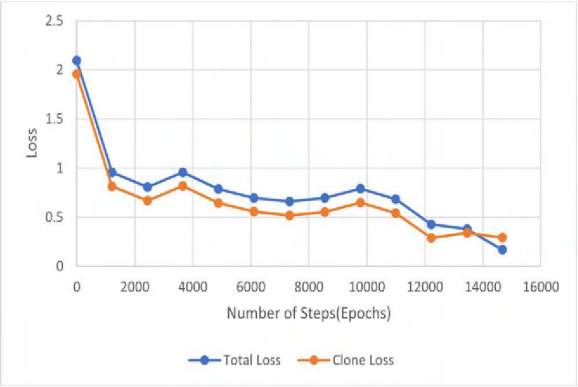

Fig. 3. Representative training loss and clone loss plots over training examples. Loss curves were selected on every epochs for display purposes.

Fig. 4 shows as the number of training epochs increases, the performance of the model is gradually increased. The model starts converging somewhere around 1500 steps, and the performance becomes stabilized when it reaches the maximum iterations 15000.

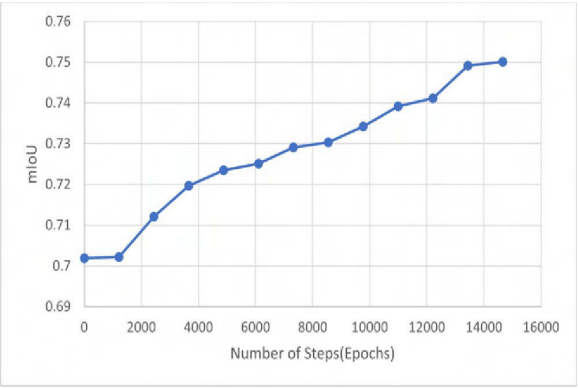

Fig. 4. Accuracy of training the model. Accuracy curves were selected on every epochs for display purposes

D. 3D Reconstruction

The segmented images can be used to build a semantically meaningful 3D model of the garbage for navigation purposes. The simplest approach is to use the original images as the input to a standard structure-from-motion or SLAM (simultaneous localization and mapping) algorithm. The semantic classification can then be stored together with the 3D coordinates of each point. Another possibility is to utilize the processed images where each semantic class represents a different color, and use these images as the input for the 3D reconstruction.

Using the color-coded images may increase the accuracy of the reconstruction, with the colored regions functioning as additional distinguishing features in the feature matching process. However, the results depend on the accuracy of the semantic segmentation: the quality of the model may deteriorate if the colors in the images indicate incorrect semantic

Fig. 5. Semantic segmentation results for material recycling Row 1: Original Image, Row 2: Test set segmented results. Indicated red box regions in which we perform better in some classes(Plastic, Wood, Sponge)

classes. It is thus necessary to determine the reliability of the segmentation network before building a 3D map from the processed images.

There are plenty of solutions available for constructing 3D models from regular perspective images, but reconstruction from 360-degree (or spherical) images is not quite as widespread despite the appearance of affordable, lightweight 360-degree cameras in the market in the last few years. A promising reconstruction technique requiring only two spherical images was presented by [15]. Such an approach can be very useful if a rough geometric map in all directions is needed for the task at hand.

The method described in [15] is based on the fact that spherical images contain photometric information from all directions, and thus the rotation between the two camera orientations can be compensated for. From the optical flow patterns between the two rotation-corrected images, it is possible to find the direction of motion of the camera via least-squares minimization. Once the camera rotation and translation (with an arbitrary metric scale) between the two images has been obtained in this manner, it remains for the 3D coordinates to be recovered via triangulation.

As in the case of perspective images, 360-degree images with color-coded semantic information can be used as an input to the reconstruction algorithm. The practical applicability of this approach is limited by the current shortage of labeled spherical image data for teaching a neural network.

III. Experimental Results

We fine-tuned the models of the Xception-65 with DeepLabV3+ networks to adapt them to the garbage semantic segmentation. We optimize the objective funtion based on weights at the network layers. We evaluate the Xception model on four videos shot in different condition at different times. The test results of the videos obtained using the trained DeepLabV3+ model, and the average mean accuracy of each class is shown in Table II.

TABLE II
Test set results on the GarbotV1 dataset, Semantic segmentation performance(mIoU) on each classes

Classes	DeepLab V3+ Backbone	mIoU
Plastic	Xception-65	79.12
Cart-board		72.13
Wood		76.19
Metal		68.98
Sponge		78.67
	Over-all	**75.01**

A. Garbage Semantic Segmentation

We employ the Xception-65 model as a backbone on DeepLabV3+ adapted for the semantic segmentation as described in section 3(b). We displayed mini-batch of 6 images which has different classes. Fig. 5 illustrates a result images in the test dataset with IoU reached 0.7501,

Fig. 6. 3D reconstruction from seven segmented frames.

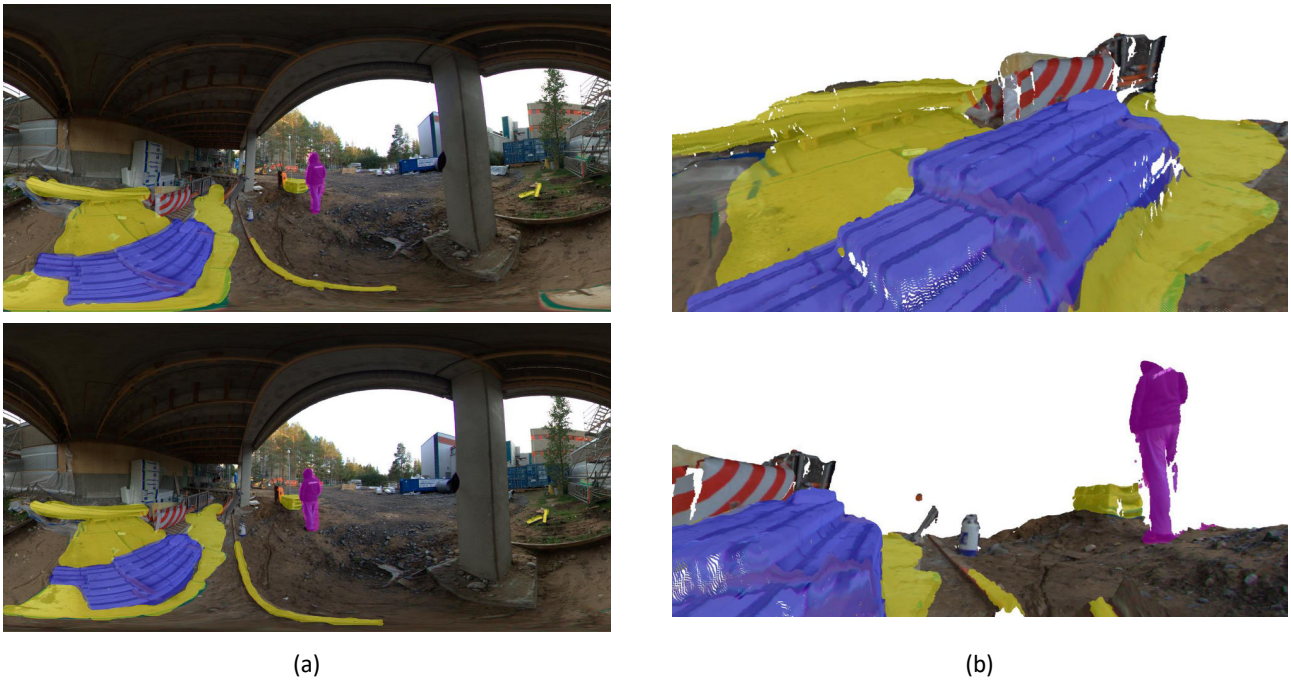

Fig. 7. 3D reconstruction from two manually segmented frames from a 360-degree camera. (a) The equirectangular projections of the spherical images. (b) Two detail images from the 3D model.

We cropped some regions from the segmented images to show our better prediction results in some of the classes such as Plastic, Wood, Sponge. It can be seen that our model achieved 76.52% on four classes and 68.98% on 1 class, showing the strong transfer repercussion and effectiveness of the model.

B. 3D Reconstruction with semantic segmentation

As an example of a 3D model constructed from our segmented images, Fig. 6 shows a reconstruction produced by the RealityCapture software ([16]) from a series of seven images with semantic classes indicated by different colors.

To demonstrate the possibilities of 3D reconstruction from 360-degree images depicting recyclable materials, we also applied the method of [15] to obtain an all-around reconstruction from two segmented images, illustrated in Fig. 7. The segmentation was done manually due to the current lack of annotated spherical image data, but the results demonstrate what the 3D models will look like in our future work. The optical flow evaluation needed to find the camera motion was obtained with the Volumetric Correspondence Network presented in [17]. This network was chosen because of its capability to handle the large distortions that appear in equirectangular images.

IV. Conclusion and Future Work

For the semantic segmentation of the garbage directly from the landfills, we developed a DeepLab V3+ network using the backbone model Xception-65, several videos of landfills were obtained for dataset and tested under varying conditions and complex hindrance factors. The important conclusions are as follows:

- The proposed method predicted and localized the different targets from the landfills on the basis of semantic segmentation. Except for certain classes, other targets have significant impact on the final results. The mean accuracy of the semantic segmentation of our method is 75.01%. This indicates the effectiveness of the segmentation directly from landfills.
- We developed a method to reconstruct the semantically meaningful 3D model from segmented images, which has huge impact on navigation of autonomous earth moving vehicles.

We will investigate the effect of implementing our method with output interfaces like earth moving vehicles and mobile robots in future work. Also, we are planning to include more significant classes into our GarbotV1 dataset.

Acknowledgment

The project 'Garbot – Utilizing robotics and AI for material recycling' was supported by the Kiertokaari Oy, Oulun Energia Oy, Finnish Environment Institute(SYKE) and Vimelco Oy. The authors appreciate them for their financial supports. The authors would also like to thank Mr. Gakwaya Achille for data annotation.

REFERENCES

[1] Kiertokaari Oy's Annual Report 2020, Finland. Available: https://kiertokaari.fi/kiertokaari-oy/vuosikertomus/ [Accessed June 01, 2021]

[2] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

[3] Kamann, C., & Rother, C. (2020). Benchmarking the robustness of semantic segmentation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8828-8838).

[4] Kaza, S., Yao, L., Bhada-Tata, P., & Van Woerden, F. (2018). What a waste 2.0: a global snapshot of solid waste management to 2050. World Bank Publications.

[5] Bobulski, J., & Kubanek, M. (2019, June). Waste classification system using image processing and convolutional neural networks. In International Work-Conference on Artificial Neural Networks (pp. 350-361). Springer, Cham.

[6] Singh, A., Aggarwal, P., & Arora, R. (2016, September). IoT based waste collection system using infrared sensors. In 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 505-509). IEEE.

[7] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857.

[8] Chen, L. C., Collins, M. D., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., ... & Shlens, J. (2018). Searching for efficient multi-scale architectures for dense image prediction. arXiv preprint arXiv:1809.04184.

[9] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

[10] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[11] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[12] Yang, M., & Thung, G. (2016). Classification of trash for recyclability status. CS229 Project Report, 2016.

[13] P. Mishra and K. Sarawadekar, "Polynomial Learning Rate Policy with Warm Restart for Deep Neural Network," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 2087-2092, doi: 10.1109/TENCON.2019.8929465.

[14] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

[15] S. Pathak, A. Moro, A. Yamashita, and H. Asama, "Dense 3D reconstruction from two spherical images via optical flow-based equirectangular epipolar rectification," IEEE International Conference on Imaging Systems and Techniques (IST), 2016, pp. 140-145.

[16] 2021 Epic Games Slovakia s.r.o, Bratislava, Slovakia. Available: https://www.capturingreality.com/ [Accessed June 01, 2021]

[17] G. Yang and D. Ramanan, "Volumetric Correspondence Networks for Optical Flow," NeurIPS, 2019.

[18] Wada, K. (2016). Labelme: Image polygonal annotation with python. GitHub repository. Available: https://github.com/wkentaro/labelme [Accessed April 29, 2021]