Automation of Waste Sorting with Deep Learning

João Sousa[1], Ana Rebelo[2], Jaime S. Cardoso[3]

***Abstract*—The importance of recycling is well known, either for environmental or economic reasons, it is impossible to escape it and the industry demands efficiency. Manual labour and traditional industrial sorting techniques are not capable of keeping up with the objectives demanded by the international community. Solutions based in computer vision techniques have the potential automate part of the waste handling tasks.**

In this paper, we propose a hierarchical deep learning approach for waste detection and classification in food trays. The proposed two-step approach retains the advantages of recent object detectors (as Faster R-CNN) and allows the classification task to be supported in higher resolution bounding boxes. Additionally, we also collect, annotate and make available to the scientific community a new dataset, named Labeled Waste in the Wild, for research and benchmark purposes. In the experimental comparison with standard deep learning approaches, the proposed hierarchical model shows better detection and classification performance.

I. Introduction

> "Action on climate change is urgent. The more we delay, the more we will pay in lives and in money"
>
> Ban Ki-Moon.[1]

We must bear in mind that global warming does not only affect humankind but also wildlife, on a very serious level. Portugal failed to meet the waste policy proposed by the European Commission[2]. The United Nations wants to ensure environmental sustainability by 2030[3]. Such targets can only be possible to achieve through an appropriate application of today's technology. On our daily lives we may forget to separate correctly the waste of our homes, and industrially speaking the companies responsible for this part have to spend a lot on labor.

Using an intelligent object identification software in waste sorting is an advantageous approach when compared to the traditional recycling methods, due to a large number of objects that are identified in a shorter period of time. The traditional approach is based on human goodwill and labor, both prone to fail on waste separation for recycling.

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project: UID/EEA/50014/2019.

[1]João Sousa is with Faculty of Engineering University of Porto, Portugal. jssousa95@hotmail.com

[2]Ana Rebelo is with INESC TEC and University Portucalense, Portugal. arebelo@inesctec.pt

[3]Jaime Cardoso is with INESC TEC and the Faculty of Engineering of the University of Porto, Portugal. jaime.cardoso@inesctec.pt

978-1-7281-5337-7/19/$31.00

Deep Learning methods are being successfully applied to diverse areas such as autonomous driving, medical imaging, and multiple industrial environments with remarkable results on object detection problems. Applying these methods to waste separation can increase the quantity of recycled material and consequently, provide an easier day-to-day life for the common person. As well as more efficacy for the industry.

In this work we address some of the identified limitations, making two main contributions: a) the proposal of a hierarchical deep model for waste detection and classification; b) a novel dataset is also presented and will be made freely available to the research community for benchmark purposes. This dataset can be used for different purposes such as a study on the types of waste produced by restaurants and the most common type of waste. To our knowledge it will be the first dataset in this field.

An overview of the existent works for classification of waste is addressed in section II followed by the description of Labeled Waste in the Wild dataset in section III. The proposed methodology is explained in section IV with the experimental study with results discussion presented in section V. Finally, conclusions are drawn and future work is outlined in section VI.

II. Related Work

Over the years, different works have been implemented with the aim of minimizing the impact of the uncontrolled disposal of waste. Maher Arebey et al. [4] propose the gray level co-occurrence matrix (GLCM) method for garbage detection and classification, combining advanced communication mechanisations with GLCM to strengthen the waste assembling operation. The proposed system uses several communication technologies including radio frequency identification (RFID), geographical information system (GIS) and general packet radio system (GPRS) integrated with a camera, streamlining the solid waste monitoring and management. The features are obtained from the GLCM and then used as inputs to a multilayer perceptron (MLP) and a K-nearest neighbor (KNN) methods for garbage separation. The results obtained show that the KNN classifier surpasses the MLP.

Sakr et al. [5] try to automatize the waste sorting by applying machine learning algorithms. The authors use two popular methods, deep learning with convolutional neural networks (CNNs) [6] and support vector machines (SVMs) [7]. The results obtained show that SVMs achieved a high classification accuracy of 94.8% while CNNs only achieved 83%.

At the TechCrunch Disrupt Hackathon, a team created AutoTrash [8], an automatic sorting trashcan that sorts garbage based on compost and recycling features. Their system has a

rotating top and uses a raspberry pi camera. The team used Google's Tensorflow AI engine and built their own layer on top of it for object recognition.

Mindy Yang et al. [9] performed a comparison study to classify waste between SVMs with scale invariant feature transform (SIFT) [10] and an eleven-layer CNN architecture similar to AlexNet [11]. Experiments show that the SVM outperforms the CNN. The accuracy obtained was 63%.

In a similar work, Oluwasanya Awe et al. [12] proposed a method using faster Region-based Convolutional Neural Networks (Faster R-CNN) [13] technique to get region proposals and object classification, reaching a mAP of 68.3%. The waste was categorized into three classes (landfill, recycling, and paper).

The authors of [14] propose an automated system based on a deep learning approach and traditional techniques to correctly separate waste into four different trash classes (glass, metal, paper and, plastic). Results showed that VGG-16[15] methods are an efficient approach for this problem, reaching 93% of accuracy on its best scenario.

The authors of RecycleNet [16] experimented on widely recognized deep convolutional neural network architectures. Training without pre-trained weights, Inception-v4 surpassed all others with 90% test accuracy.The authors then applied transfer learning and fine-tuning of weight parameters using ImageNet weights, and DenseNet121 obtained the best result with 95% test accuracy.This latter solution has slower prediction time.

In a nutshell, in the last years, computer vision has being considered as a tool to support waste sorting, and deep learning techniques have reached reasonable results in controlled scenarios. Object detection, less addressed in waste management, was considered in Oluwasanya Awe et al [12] using the Faster R-CNN model, showing reasonable results.

Although hierarchical classification has been explored in other domains [17] to bring robustness and efficiency, it has not yet been explored in waste sorting.

III. Labeled Waste in the Wild dataset

Although several waste datasets have been proposed in the literature, many issues remain unexplored: (i) there are no datasets available that contain different types of waste in the same image; (ii) most of the available datasets were recorded in a very controlled scenario; (iii) there are no waste datasets that were acquired with different sensors. The dataset presented in this paper attempts to address all of these problems.

The Labeled Waste in the Wild dataset contains 1002 images of used food trays using several different smartphones. Depending on the application, a real-world setting has a significant weight on the performance of a model. In this manner, our dataset is composed by food tray images acquired from different shopping centers, canteens or home, with no control over lighting, objects position and type.

The images are 3456×4608, RGB, taken inside building, with good artificial light conditions. We enlarged our in-house dataset with 180 images from the Ciocca et al. [18] food dataset. Afterwards, images were annotated on Labelbox.com. Illustrative examples are included in Fig. 1.

(a) Example 1. (b) Example 2. (c) Example 3.

Fig. 1. Illustrative images and respective annotations.

The dataset is available upon request to the authors.

A. Classes and Multi-label perspective

The dataset is composed of 19 different classes totaling 7200 labels, all classes represent a shape and the material they contain as we can see in Table I.

TABLE I
FREQUENCY OF EACH CLASS IN THE DATASET.

bottle_glass	26	bottle_plastic	266
box_paper	255	box_plastic	54
can_metal	112	cup_glass	185
cup_paper	423	cup_plastic	225
cutlery_metal	703	cutlery_plastic	114
mixed_waste	249	paper_napkin	1086
plastic_waste	207	plate_ceramic	932
plate_glass	18	plate_plastic	124
plastic_straw	173	tray	997
paper_paper	1067	total	7216

In a real-world application, not every object is considered waste. For that reason in this dataset a significant portion of the objects are not waste. It is important that our model learns to distinguish variety. Regarding types of waste, the difference between the class paper_paper and paper_napkin is the presence of organic waste in it, a piece of relevant information for industrial recycling.

Multi-label is a characteristic of this annotation because the labels can be rearranged in two groups of classes, either by shape or material, and we use that to our advantage in the object detection and classification. The annotation can be framed in a multi-label setting, where each object is annotated in respect to the shape and to the material (and the bounding box). The material label can take values in the set {glass, paper, metal, plastic}. The shape label can take values in the set {cup, plate, box, tray, cutlery, mixed_waste, bottle, paper, can, plastic}.

B. Cross-Sensor Application

Labeled Waste in the Wild contains 400 more images for cross-sensor validation of algorithms. The images are

composed of photos of two different sets, each containing 100 used food trays. Each set has two images of the same tray using different smartphone cameras. The first set was obtained using an Honor 10 and a Xiaomi Redmi Note 3, and the second set was obtained using the Honor 10 and Iphone 5. The images were labeled in the same conditions as the original 1002 images of the dataset.

IV. Hierarchical Deep Detection and Classification of Waste

Consider a set of images of waste in food trays. The goal is, in each image, to detect the bounding boxes of all the waste objects present, and assign each detected object to one of the considered classes.

A. Standard baseline approach: flat method

The standard baseline approach consists on building a model that uses all images and detects for each input simultaneously the bounding box and the waste class. This approach is the typical deep learning approach in which all efforts are supported by the neural network and has usually good results. We adopted the state-of-the-art algorithm Faster R-CNN [13] in object detection and classification.

Since Faster R-CNN (and similar competitors) resizes all images to a fixed, low spatial resolution (600 pixels), for computational efficiency, a lot of information is lost when resizing the images from their 3456×4608 original size.

1) Improved baseline approach: Cascaded Region Proposal and Classification: To overcome the spatial resolution issue, in this method Faster R-CNN is again trained in the flat semantic label space of 19 classes, but we only keep the proposed bounding boxes. In a second step, using the Bounding box coordinates in the original image, the original image is cropped and resized to 500×500, and fed as input to a standard CNN for (flat) classification. As seen in Fig. 2.

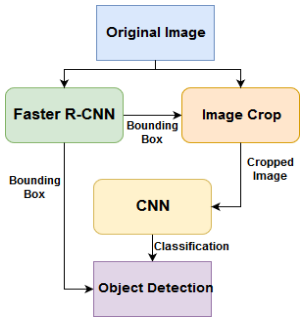

Fig. 2. Cascaded Region Proposal and Classification

The adopted CNN is quite conventional, with just as few layers, as depicted in Fig. 3. The advantage of this strategy is that this CNN processes images with more detail than present in Faster R-CNN, enabling better classification accuracy.

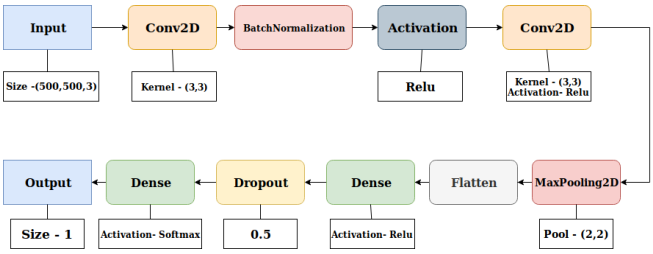

Fig. 3. CNN Architecture

B. Hierarchical Approach

The rationale for the proposed two-step hierarchical deep learning approach is to divide the semantic classification problem in two subproblems. In a first step, Faster R-CNN is used as informative region proposal, we keep the bounding box and the classification, a mother-class based on shape or material, taking advantage of the multi-label characteristic of the LWW dataset as mentioned in Section III-A. In a second step, the bounding box obtained in the original image is cropped, resized to 500×500 and fed to the CNN, with the same architecture detailed in Fig. 3, trained for the specific mother-class, outputting a child-class based on shape or material. Splitting the hierarchical approach into shape and material was inspired by the work published by Geirhos et al [19].

1) Material Based: As aforementioned, Faster R-CNN is trained for 4 mother-classes based on material{glass, paper, metal, plastic}. Four CNNs are trained, one for each mother-class, to prevent the detection of a class that does not exist. For example, if the mother class is metal then we know that the child class cannot be a cup.

This model receives an image as input, then the Faster R-CNN outputs the bounding box and the mother-class. With that information, the original image is cropped, which is the input of the CNN related to the mother-class. Then outputs the child-class, that merged with the mother-class, forms the final class. The latter, with the bounding box predicted before forms the detection of the model. The architecture of this model can be seen in Fig. 4.

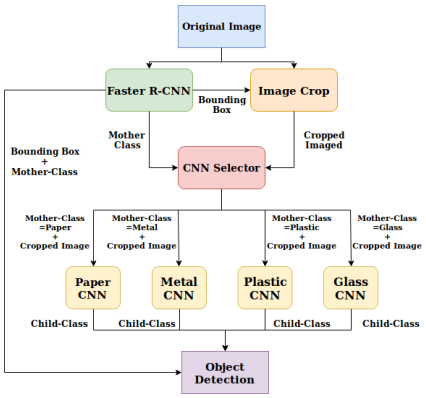

Fig. 4. Hierarchical based on material architecture

2) Shape Based: In this method Faster R-CNN is trained for 10 mother-classes based on their shape{cup, plate, box, tray, cutlery, mixed_waste, bottle, paper, can, plastic}. Some mother-classes are unique and do not need a CNN associated with them, if the output of the Faster R-CNN is tray, mixed_waste or can, then they are the final class.

Seven CNNs are trained for the same reason mentioned in IV-B1, this time if the mother-class is a cup then the child class cannot be metal.

This model behaves very similarly to the material based one as detailed in Fig. 5, an image is used as input to the Faster R-CNN that outputs a bounding box and a mother-class. In this case it is a shape. An image, result of the cropped original image using the bounding box coordinates, is the input of the CNN. This outputs a child class, a material, and together the final class and bounding box correspond to the detection of the model.

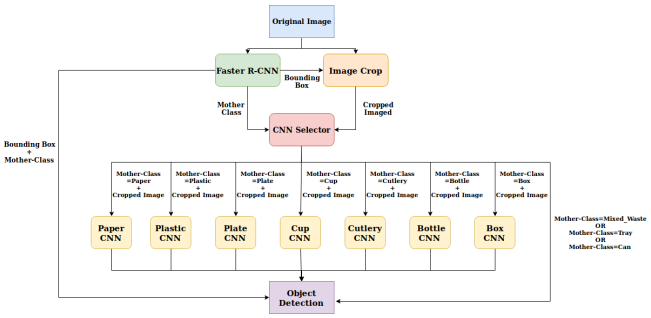

Fig. 5. Hierarchical based on shape architecture

V. EXPERIMENTAL STUDY

In the experiments the same model of Faster-RCNN was used, from the LWW dataset 800 images were used for training, for 100 epochs with a duration of 40 hours, and 200 images for testing. Regarding the CNN training, some data augmentation techniques were used, such as changing brightness and flipping the images, in order to increase the size of the training dataset.

The metric of evaluation used was the mean average precision (mAP), it is the method used to evaluate the state-of-the-art algorithms in object detection and classification [13].

A. Flat method

The direct use of Faster R-CNN model, trained in the 19 original classes, obtained an mAP of 74.1%.

Cascaded Region Proposal and Classification: The cropped images obtained by the ground truth bounding boxes of the LWW dataset were used for training the CNN. A total of 7216 images were used: 6400 for training and 800 for validation, chosen randomly. The best accuracy obtained with data augmentation was 67% and after integration with the Faster R-CNN, the mAP was 81.4%.

B. Hierarchical

Material Based: Faster R-CNN was trained for 4 mother-classes and the mAP obtained before integration with the CNNs was 72.8%.

The training process of the CNN was similar to the previous method. However they were split by material and used for the mother-class specific CNN, the child-classes were based on shape. The results obtained were the following:

- CNN_paper: trained for 4 child-classes with 2272 images for training and 560 for validation reaching an accuracy of 42%;
- CNN_plastic: trained for 9 child-classes with 1920 images for training and 480 for validation reaching an accuracy of 67%;
- CNN_glass: trained for 4 child-classes with 928 images for training and 224 for validation reaching an accuracy of 78%;
- CNN_metal: trained for 2 child-classes with 640 images for training and 160 for validation reaching an accuracy of 84%.

After integration with the Faster R-CNN, the mAP obtained was 80.9%.

Shape Based: Faster R-CNN was trained for 10 mother-classes the mAP obtained before integration was 73.6%.

The training process of the CNN was similar to the previous method. The results obtained were the following:

- CNN_bottle: trained for 2 child classes with 224 images for training and 64 for validation reaching an accuracy of 90%;
- CNN_box: trained for 2 child classes with 240 images for training and 64 for validation reaching an accuracy of 84%;
- CNN_cup: trained for 3 child classes with 672 images for training and 160 for validation reaching an accuracy of 72%;
- CNN_cutlery: trained for 2 child classes with 656 images for training and 160 for validation reaching an accuracy of 89%;
- CNN_paper: trained for 2 child classes with 1728 images for training and 416 for validation reaching an accuracy of 75%;
- CNN_plastic: trained for 2 child classes with 288 images for training and 80 for validation reaching an accuracy of 73%;
- CNN_metal: CNN_plate was trained for 3 child classes with 832 images for training and 240 for validation reaching an accuracy of 87%.

After integration with the Faster R-CNN, the mAP obtained was 86%.

C. Cross-Sensor Validation

Out of all the models tested, the hierarchical method based on shape achieves the best performance. The 200 images of the LWW dataset used to test the models were obtained using the same camera (Honor 10 smartphone). The cross-sensor

validation method consists in testing the same model on images of the same trays obtained with different cameras, hence the name "cross-sensor". The LWW dataset was designed to enable cross-sensor validation, apart from the 1002 images, 400 more images were added. From those 400 images, two different sets of 100 trays are present, from each set, two different cameras took the photos, resulting in 200 images of each set. From the first set, an Honor 10 and a Xiaomi Redmi Note 3 were used, and in the second set, an Honor 10 and an Iphone 5 were used. The mAP score of the hierarchical shape based model are present in Table. II.

TABLE II
CROSS-SENSOR VALIDATION MAP SCORE

Set	sensor	mAP(%)
1	Honor 10	90.7
1	Xiaomi Redmi Note 3	92.9
2	Honor 10	88.6
2	Iphone 5	84.0

Each set has the same trays and objects, the only difference being in the sensor. Having this information into account, results show that the sensor has influence on the performance of the algorithm. Honor 10, the most used sensor on the LWW dataset has a worse perfomance than Xiaomi Redmi Note 3, but a better performance than Iphone 5. The fact that the mAP is calculated on 100 images has the disadvantage of some classes having only a few representations in the dataset, influencing the individual AP of that class and ultimately, influencing the mAP score.

D. Discussion

The results of our methods are summarized in Table III.

TABLE III
MODEL RESULTS

Model	mAP (%)
flat	74.1
Cascaded	81.4
Hierarchical - Material	80.9
Hierarchical - Shape	**86**

Differences are noticeable when comparing the flat and cascaded approach, the latter improved the mAP in 7.3%.

Faster R-CNN resizes the images to a maximum of 600 pixels on height or width, that causes loss of information on our 3456×4608 images. This improves the duration of training and prediction with a good performance when the classes have significant differences. Moreover, there is a huge difference between an airplane, a car and a person, however that is not the case between a cup of paper, a cup of glass or a cup of plastic. Intuitively, if an image has a lower resolution some details will be lost, making the classification of Faster R-CNN harder. In order to overcome this limitation, our CNN uses the bounding box coordinates predicted by Faster R-CNN re-scaled to the original size, proving improved results on complex classes.

The hierarchical method based on shape obtained the best result, compared to the flat approach, improved the mAP in 11.9%. In this case it is a matter of complexity. Faster R-CNN does not get significant mAP changes with the number of classes as shown in Table IV.

TABLE IV
MOTHER CLASS RESULTS

Model	mAP (%)
Faster R-CNN flat	74.1
Faster R-CNN - Material	72.8
Faster R-CNN - Shape	73.6

However, the number of classes is significant on the CNNs. Each CNN on the material based method has more classes to predict (for instance CNN_plastic was trained for 9 child classes) whereas in the shape based method most of the CNNs were trained only for two child classes. Hence the difference of 5.1% in the mAP score.

VI. CONCLUSION AND FUTURE WORK

This paper presents a novel LWW dataset and an improvement on the Faster R-CNN in order to perform an end-to-end waste detection and classification. As future work, we plan to increase our LWW dataset with the purpose of creating even more variability to test it with different CNNs architectures and perform cross-sensor validation approaches. Another research line is the application of our proposed model on video, using a robotic system for waste sorting.

REFERENCES

[1] (Sep. 16, 2014). Ahead of general assembly kick-off, un chief urges world leaders to unite in 'time of turmoil' — africa renewal online, [Online]. Available: https://www.un.org/africarenewal/africaga/news/ahead-general-assembly-kick-un-chief-urges-world-leaders-unite-'time-turmoil' (visited on 02/05/2019).

[2] A. Coentrão, "Comissão europeia põe em causa números da reciclagem em portugal," *Público*, Oct. 8, 2018.

[3] U. Nations, "About the sustainable development goals," *un.org*,

[4] M. Arebey, M. Hannan, R. A. Begum, and H. Basri, "Solid waste bin level detection using gray level co-occurrence matrix feature extraction approach," *Journal of environmental management*, vol. 104, pp. 9–18, 2012.

[5] G. E. Sakr, M. Mokbel, A. Darwich, M. N. Khneisser, and A. Hadi, "Comparing deep learning and support vector machines for autonomous waste sorting," in *Multidisciplinary Conference on Engineering Technology (IMCET), IEEE International*, IEEE, 2016, pp. 207–212.

[6] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazare, T. Bui, and J. Collomosse, "Everything You Wanted to Know about Deep Learning for Computer Vision but Were Afraid to Ask," *Proceedings - 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials SIBGRAPI-T 2017*, vol. 2018-Janua, pp. 17–41, 2018. DOI: 10.1109/SIBGRAPI-T.2017.12.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] J. Donovan, *Auto-trash sorts garbage automatically at the techcrunch disrupt hackathon*, 2016.

[9] G. T. M. Yang and G. Thung, "Classification of trash for recyclability status," *CS229 Project Report*, vol. 2016, 2016.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] O. Awe, R. Mengistu, and V. Sreedhar, "Smart trash net: Waste localization and classification," 2017.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017, ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2577031. arXiv: 1506.01497.

[14] B. S. Costa, A. C. S. Bernardes, J. V. A. Pereira, V. H. Zampa, V. A. Pereira, G. F. Matos, E. A. Soares, C. L. Soares, and A. Silva, "Artificial intelligence in automated sorting in trash recycling," Oct. 2018, pp. 198–205. DOI: 10.5753/eniac.2018.4416.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] C. Bircanoğlu, M. Atay, F. Beşer, Ö. Genç, and M. A. Kızrak, "Recyclenet: Intelligent waste sorting using deep neural networks," in *2018 Innovations in Intelligent Systems and Applications (INISTA)*, IEEE, 2018, pp. 1–7.

[17] I. Amaral, F. Coelho, J. F. P. da Costa, and J. S. Cardoso, "Hierarchical medical image annotation using svm-based approaches," in *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine*, 2010.

[18] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition: A new dataset, experiments and results," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 588–598, 2017. DOI: 10.1109/JBHI.2016.2636441.

[19] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.