Synthetic data generation based on grid deformation for waste recycling applications

Nick Tsagarakis
Institute of Computer Science
FORTH
Heraklion, Greece
nikolats@ics.forth.gr

Alexandros Antonaras
University of Nicosia
School of Business
Nicosia, Cyprus
antonaras.a@unic.ac.cy

Michail Maniadakis
Institute of Computer Science
FORTH
Heraklion, Greece
mmaniada@ics.forth.gr

***Abstract*—Today, waste recycling is supported by intelligent robots that use machine learning to identify and sort recyclables. The development of computer vision applications based on machine learning relies heavily on large datasets that are used to train deep neural network models. In recent years, methods that allow the creation of large training datasets from a limited initial set of images have been investigated. This paper describes a method in which segmented images of real recyclables (polyethylene terephthalate, PETE) are artificially deformed using mesh transformation to create new instances of the recyclable objects. The new instances are placed on real backgrounds to create synthetic images. This process allows the generation of large artificial datasets used for training neural networks. We evaluate the usability of these datasets by studying the extent to which they can improve the performance of trained models when applied in real and challenging industrial images. In particular, we consider the main metrics used to evaluate the performance of classification models, namely Accuracy, Precision and Recall. The results obtained show that including even small-scale object deformations in the artificial datasets can slightly improve the Accuracy and significantly improve the model Recall, while Precision of the model remains unchanged.**

***Index Terms*—Synthetic waste data, Grid deformation, Deep Learning, Computer Vision, Material recovery**

I. Introduction

The recovery and recycling of post-consumer packaging materials is central to the circular economy model that has become increasingly popular in recent years. The term thermoplastics refers to a group of materials that are widely used in packaging and are suitable for recycling. Due to their mechanical and chemical properties, thermoplastics can be melted and recast many times and at the same time retain all their usability properties. One of the best thermoplastics in terms of recyclability and processability is polyethylene terephthalate, commonly known as PETE or PET. This material is a primary target for Material Recovery Facilities (MRFs).

The last years, robots guided by advanced computer vision systems have been installed in MRFs to speed up material recovery and enhance the processing capacity of such facilities. The real-time detection and categorization of recyclables as they are transported on an industrial conveyor belt is a particularly challenging task for computer vision units.

To date, the dominant approach regards the use of deep neural networks which are trained to identify and categorize recyclables. The training of deep neural networks is based on large datasets consisting of images depicting the recyclable waste transported on the industrial belt. However, the annotation of the *"belt-images"* used for training is a difficult and very time-consuming process that requires many resources to be devoted to manual image processing.

To overcome the need for extensive manual image annotation, we investigate synthetic data generation methods that allow the acquisition of large, already annotated datasets. We focus on implementing an easy-to-use, low-cost method, which can generate a large number of new synthetic images that increase variance in the features of the dataset. The use of the new, enhanced dataset for training deep NNs is expected to improve the efficiency of the obtained model in comparison to the one trained with the original limited-size dataset.

The present work focuses on a key challenge for MRFs that is the categorization and separation of materials coming in mixed recyclable waste streams. In particular, we focus on the identification of PETE, a plastic widely used in water and juice bottles, which currently has one of the highest prices in the secondary materials market. Due to the latter, MRFs spend a lot of resources in PETE collection. Our work aims to automate and improve computer vision based PETE identification in mixed waste streams (this aims to be integrated with the robotic PETE picker we have implemented in previous works [1]–[3] (see the following Youtube-link).

The computer vision module explored in the present study is based on the well known Mask R-CNN deep neural network that is trained to identify PETE recyclables. We start with a dataset containing 1000 industrial images from the conveyor belt of an MRF, where PETE objects are manually annotated. For the rest of the paper this is called the *"Base"* dataset.

Additionally we collect 440 images of segmented PETE objects which will be used to generate synthetic datasets. We implement a method that computationally deforms the segmented and isolated *"object-images"* to create multiple slightly different instance of the object after applying random grid based image distortions. The new, artificial objects-images

This work is co-financed by the European Union's Horizon Europe Research and Innovation program under project RECLAIM GA: 101070524, and additionally by the European Union and national resources of Greece and Cyprus within the framework of the INTERREG V-A Cooperation Program "Greece – Cyprus" 2014-2020, project InterRecycle MIS: 5047863.

are superimposed on randomly selected belt-images to create a new dataset. We generate four different synthetic datasets, consisting of artificial PETE objects developed from varying degrees of mesh deformation. The synthetic datasets are used for training Mask R-CNN models, which are further evaluated on the recognition of unseen PETE objects.

In short, this paper aims to (i) present an easy to develop grid-deformation approach for generating synthetic recyclable waste data (ii) contrast the usability of manually annotated and artificially generated datasets in training Mask R-CNN based computer vision models for waste sorting, (iii) examine the extent of deformations that are necessary for having a positive effect on the machine learning potential of the synthetic dataset.

The current paper is organized as follows. Section 2 provides a short review of the synthetic data generation literature. Then section 3 discusses the proposed method providing several implementation details. The next section presents the results obtain after applying the proposed method to create a PETE identification module. Section 4 provides a discussion on the obtained results. Finally, the last section provides conclusions and direction for fruitful future work.

II. Literature Review

In recent years, deep learning has been used to tackle many real-world problems, including the identification and categorisation of recyclable waste [1], [2], [4]. Deep learning assumes the use of large, well defined datasets to train multi-layer neural network models. However, to acquire large datasets, extensive real-world recording is required. This task is not easy, and will then have to be complemented by laborious manual work to annotate the recorded data. Overall, data logging and annotation is currently considered a key hurdle for machine learning applications [5].

An alternative approach, regards the collection of much smaller scale data which are then used as a starting point for creating new, realistic, synthetic data that are included in the training of the deep neural networks [6], [7]. Following this approach one can not only save significant time but also minimize costs related to data collection and annotation [8].

In computer vision applications, it is often easy to record a high number of images. However after that, it is necessary to manually process these images to label regions of interest with pixel-level precision. The latter makes the acquisition of large well-annotated image datasets particularly challenging. Therefore, the generation of synthetic image data is particularly relevant for many computer vision applications [9], [10]. State of the art approaches rely on Generative Adversarial Networks [11], Variational Autoencoders [12], Diffusion Models [13], and other task specific networks. However these approaches assume again the collection of data for training the generative models that produce the synthetic images.

In contrast to the above, the current work considers the use of a straightforward algorithmic method for the generation of synthetic recyclable waste data. The proposed approach is particularly low-cost and assumes only a small amount of annotated data to be available for the generation of synthetic data sets.

Fig. 1: A sample of object-images extracted from industrial belt-images.

Fig. 2: An exemplar images taken from the belt conveying waste in the material recovery facility.

III. Methodology

As discussed above, the present work considers the advancement of existing datasets by using available segmented *object-images* (see Fig 1) which are artificially deformed to be superimposed over real-world *belt-images* that depict the flow of waste on a conveyor belt in the industrial context of a MRF (see Fig 2). The goal is to develop a richer, automatically generated dataset that, when used for training, improves the generalization of the obtained model, in comparison to the original dataset.

To achieve this goal, we have developed a two-step process that first takes object-images to generate multiple, new randomly deformed object-images. Then, it places the new images on top of belt-images and updates the relevant annotation file to generate an advanced synthetic dataset. These two steps are described in the following sections.

It is important for our work to develop a PETE identification module that is applicable in real industrial environments. The recording of data is performed in the material recovery facility of Heraklion, Crete, Greece.

A dark room is implemented above the belt conveying waste to minimize the effect of external lighting. Inside the dark room we install a 2MP RGB camera and three light bars providing uniform ambient light on the waste. We use this installation to record one image every second, for several

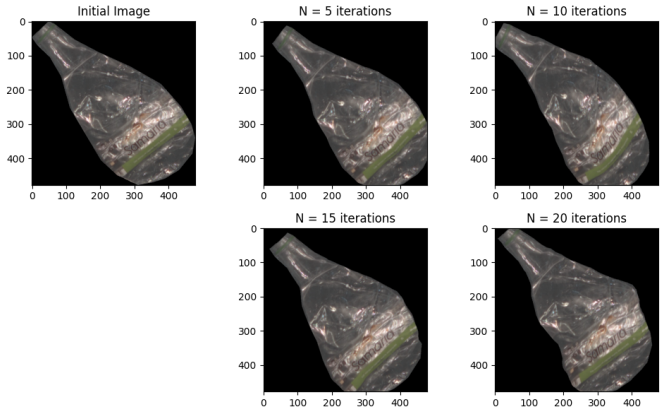

Fig. 3: An initial object-image (left) undergo 5, 10, 15 and 20 distortions (right).

hours. Following this approach we have been able to obtain more than 10K belt-images recorded in a real and particularly difficult industrial context. On average, each image contains 3.2 PETE objects.

We use 1000 randomly selected images to be manually annotated using the online tool: VGG Image Annotator [14]. This set of images composes the so called "Base" dataset. Additionally, a json file is created, to provide a meta-description of the content of the Base dataset.

A. Random deformation of segmented object-images

To develop a large number of synthetic new object-images using a limited set of initial object-images we use grid deformation that can easily and effectively generate geometric transformations on images. This is implemented using the inverse distance weighting interpolation [15], assuming that the displacement of a grid point is propagated to the interior of only the surrounding grid. The weighting function that controls the appearance within the grid is directly related to the distance between the moved interior point and boundary points.

Totally, 440 different manually isolated object-images are used for data generation. The original PETE object images are considered as rigid bodies and under the stresses that undergo in the real world, the deformation that is created at one point will propagate only to the neighbour points. It is assumed, there are two kinds of deformations: folds and curves.

We create a 2D mesh with dimensions m x n, which is applied on every given object-image of dimensions m x n. A random node of that grid is selected to move by a random generated vector. The magnitude of that vector, which describes the extend of that movement, takes a number in the range between 10 and 30 pixels. The orientation of the vector is within 0 and 360 degrees. Then, accordingly to an inverse distance weighting interpolation, every other node of the mesh is moved, simulating a rigid body distortion.

The number of times that the above procedure takes place is denoted as N and defines the shape of final deformed mesh that will be applied to a real object-image. We interpolate it linearly at pixel level. Essentially, every iteration of that process, inserts a fold or a curve to the object-image, simulating the physical distortion that a waste object may undergo.

B. Synthetic Dataset Generation

Synthetic datasets consist of artificial belt-images, that are created by superimposing the deformed objects-images on top of industrial, originally PETE-free, belt-images.

We use 440 real PETE object-images which are randomly deformed ten times to come up with 4400 new artificial object-images. These are placed over 500 real belt-images used as backgrounds (see Fig 2). On each background image, we superimpose a random number (between 1 and 5) of artificial object-images to generate a new synthetic belt-image (see Fig 4). The synthetic belt-image is annotated automatically, by entering the border line of the added objects in the *.json* file describing the content of the dataset.

Following the above, we generate four synthetic datasets each one generated by using a different set of deformed object-images. In particular, the 5Def synthetic dataset is implemented by using 5 deformation steps, the 10Def synthetic dataset is implemented by using 10 deformation steps, the 15Def synthetic dataset is implemented by using 15 deformation steps) and finally, the 20Def synthetic dataset is implemented by using 20 deformation steps.

Fig. 4: A synthetic image generated by superimposing 5 PETE bottles on top of the background image shown in Fig 2.

IV. Results

To evaluate the synthetic dataset generation procedure we examine the ability of the generated datasets to enhance the learning capabilities of the deep convolutional neural network Mask R-CNN.

In particular, we examine the quality of Mask R-CNN training when using the base dataset, and the 5Def, 10Def, 15Def, 20Def synthetic datasets. Furthermore, the synthetic data generation method is contrasted to the standard augmentation approach that is frequently applied on deep neural network training datasets and includes standard random affine, scale and rotation transformations, hue, saturation and colour modifications and crop augmentation.

A. Model Training

As discussed in previous works [2] waste identification in an industrial context assumes the solution of the problem known as "instance segmentation". This is necessary because the model needs to identify and categorize multiple, potentially overlapping objects that are carried over the industrial belt. The well-known Mask Regional CNN (Mask R-CNN) [16] has been widely used in the past to successfully address real world instance segmentation tasks and is therefore adopted in the present work to develop the PETE identification module.

It is noted that in all experiments we use the same training parameters (in fact they have been specified through trial and error procedure, to improve the performance of the base model). The relevant parameters are listed bellow.

- backbone network: Resnet101
- image resize mode: 512 x 512 px.
- batch size: 12 RGB images
- learning rate : 0.002
- number of epochs : 100
- steps per epoch: 500 (for datasets of 10000 images) and 150 (for base dataset of 1000 images)

Moreover, Mask R-CNN uses a multi-objective loss function which is calculated as the weighted sum of different losses (*Rpn class loss*: presence or absence of objects, *Rpn bbox loss*: area bounding box, *Mrcnn class loss*: categorization of the object, *Mrcnn bbox loss*: object bounding box, *Mrcnn mask loss*: pixel-level object localization). It is therefore important to specify the weights of the individual losses in estimating the global loss values. For computer automated solutions applied to the recovery of recyclable materials, it is important that (i) the objects are identified, (ii) the objects are categorized in the correct class to avoid mixing material types and (iii) the boundaries between objects are correctly specified to facilitate correct estimation of the objects' centroid.

Therefore during Mask R-CNN training, we provide higher weights to the partial loss functions addressing the above issues. Specifically, the weights of the individual criteria used for training the Mask R-CNN are set as follows:

- rpn class loss: 3
- rpn bbox loss: 1
- mrcnn class loss: 3
- mrcnn bbox loss: 1
- mrcnn mask loss: 4

The above mentioned training parameters have been used in all experiments discussed in the present work.

B. Testing

We train six different Mask R-CNN models using five different datasets. The first model is trained using the base dataset. The second model is trained using the same base dataset, but this time with the random image data augmentation activated. The rest four models are trained using the 5Def, 10Def, 15Def, 20Def synthetic datasets, without any image augmentation during model training.

Fig. 5: An example of model prediction on a first time seen image, used for testing.

To contrast the performance of the models we use a set of 200 real images, unseen during training. To facilitate evaluation, we have manually specified the groundtruth for all testing images. We use the six trained models mentioned above, to identify PETE objects in the testing images (see Figue 5). After testing each model we create the corresponding confusion matrix, by which we calculate the values of the fundamental metrics of the Mask R-CNN classifier at pixel level, namely model Accuracy, Precision and Recall.

Accuracy is defined as the percentage of overall correct classified pixels that a model is able to predict. Due to the imbalance of data, which means that PETE objects occupy substantially less space than prevalent background, we also calculate the so called Balanced Accuracy which considers the fact that background appears much more that PETE. Precision describes the percentage of correct predictions for one class (for PET, in current occasion) relative to all predictions. Recall refers to the percentage of actual PETE pixels that the model is capable of identifying.

According to the PETE identification results summarized in Table I, the models trained with the synthetic data developed with a predefined number of deformation iterations outperform the base model. In particular, the accuracy of the Base model is lower than the rest. The augmentation of the Base dataset images improves model accuracy. However further improvement is achieved when the synthetic images are used for training instead of the original dataset. The difference in model performance is further revealed after considering the balanced accuracy of model predictions.

According to Table I, the precision of all models have similar values. However, the models trained with synthetic data have higher Recall ability, which means that these models miss less PETE objects depicted in the test images. The model trained with the data containing the highest extent of deformations (i.e. Def20) exhibit the highest Recall ability.

Elaborating further on Recall results, we note that the augmentation of the Base dataset has also improved performance compared to the case of using the plain Base dataset. Still, there is further improvement when the synthetic datasets with

Training Dataset	Accuracy	Balanced Accuracy	Precision	Recall
Base	97.34 %	86.21 %	90.71 %	72.65 %
Base + Augmentation	97.88 %	88.70 %	91.67 %	77.87 %
Def5	98.27 %	90.41 %	90.17 %	81.27 %
Def10	98.18 %	90.60 %	90.22 %	81.73 %
Def15	98.25 %	90.51 %	91.75 %	81.43 %
Def20	98.23 %	91.12 %	90.06 %	83.67 %

TABLE I: The performance of the Mask R-CNN models trained with different datasets.

Training Dataset	Mean Jaccard coefficient
Base	70.61 %
Base + Augmentation	74.16 %
Def5	77.61 %
Def10	77.31 %
Def15	77.85 %
Def20	78.35 %

TABLE II: The effect of the number of deformation iterations on Mask R-CNN model performance.

artificially deformed PETE objects is used for training. We believe this is due to the fact that the image augmentation techniques (e.g. saturation) are applied universally to the whole extent of an image, while in contrast the generation of synthetic data based on object deformations has a local effect trying to provide more instance of how PETE objects may look-like. It is necessary to note here that we have run training sessions with the synthetic datasets and the image augmentation activated, which surprisingly have not improved the performance of the generated models any further.

Complementary to the above, we visually inspect the accuracy of the "masked" PETE areas inferred by the individual models. This is illustrated in Figure 6, which shows the performance of the model trained with the base dataset and the ones trained with synthetic data. The accuracy of the borders of the identified objects are clearly more accurate in the images that correspond to 5Def and 10Def training dataset, which correspond to the N=5 and N=10 object-image deformations.

Besides visually inspecting the difference between the inferred object area and the ground-truth, we quantify their similarity, which provides an additional aggregated and comparable measure of the performance of each Mask R-CNN model. To this end, we consider Jaccard coefficient that contrasts the inferred against the ground-truth region using the ratio of the intersection over the union of the two regions. The averages of the Jaccard coefficient measuring the similarity of the actual and predicted regions over 200 test images are summarized in Table II. According to these results, the models trained with the synthetic datasets outperform the base model.

Overall, the results summarized in Tables I and II show that the synthetic datasets automatically created using artificially deformed object-images can lead to better models compared to the use of regular-size dataset annotated by humans (after many demanding work hours). Interestingly, the use of a large number of artificially generated data outperforms the model performance even when random image augmentation is adopted during the training of the model.

V. Discussion

According to the results summarized in the previous section, there is a significant improvement in the performance of the Mask R-CNN model when it is trained using synthetic datasets generated by the elastic deformation of images to achieve the creation of new visual instances of the target objects.

The use of synthetic data slightly improves the Accuracy of the trained models, while their Precision remains at the same level. What significantly improves is the Recall ability of the models. In particular, the Mask R-CNN models trained with the use of synthetic data tend to achieve much higher recall rates, since their ability to identify PETE objects in real data is skewer than the base model's.

This improvement is observed by the use of all the synthetic datasets created in the present work for Mask R-CNN training, regardless of the extent of object deformations. To further assess the role of object-image deformations, we have also generated a new dataset by simply copying and pasting object-images but without any kind of deformation, on top of the background images. The use of this dataset for training has given very similar results to the Base dataset. Therefore, we conclude that the deformation of object-images are the main reason for improving the performance of Mask R-CNN models.

VI. Conclusions

The current work investigates the usability of synthetic datasets in training Mask R-CNN models. The datasets are created by the random, artificial deformation of object-images which are placed on top of background images to create large number of automatically annotated images.

The use of deformed object-images is shown to improve the overall performance of the trained models. In particular, training with synthetic data results into slight improvement in model's Accuracy, same performance in terms of Precision and significant improvement in terms of the Recall abilities of the trained model.

The results of the present work will provide the basis for the development of more synthetic datasets that will concern other recyclable materials, such as aluminum packaging and HDPE plastic packaging. We aim at the generation of extended multi-material datasets which will allow the training of Mask R-CNN models that will recognise multiple high-value recyclables in challenging conditions. In addition, we are interested in developing recursive pipelines that use the images of objects correctly identified by the current solution to generate a new improved dataset, which is used to develop a better trained model. The improved model can be further applied to new images to identify new objects that can again be incorporated into the dataset and so on.

Moreover, our ongoing work is focused on integrating the Mask R-CNN based recyclable detection and categorization modules with the robotic systems we have been developing in recent years for sorting recyclables in high demand industrial conditions.

Fig. 6: Indicative identification of PETE bottles by the models trained with the different datasets considered in the current work. All images are zoomed to facilitate visual inspection. In all images model prediction polygon is shown in orange and groundtruth polygon is shown in blue.

REFERENCES

[1] K. Moirogiorgou, F. Raptopoulos, G. Livanos, S. Orfanoudakis, M. Papadogiorgaki, M. Zervakis, and M. Maniadakis, "Intelligent robotic system for urban waste recycling," in *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2022, pp. 1–6.

[2] M. Koskinopoulou, F. Raptopoulos, G. Papadopoulos, N. Mavrakis, and M. Maniadakis, "Robotic waste sorting technology: Toward a vision-based categorization system for the industrial robotic separation of recyclable waste," *IEEE Robotics & Automation Magazine*, vol. 28, no. 2, pp. 50–60, 2021.

[3] F. Raptopoulos, M. Koskinopoulou, and M. Maniadakis, "Robotic pick-and-toss facilitates urban waste sorting," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 2020, pp. 1149–1154.

[4] F. Shennib and K. Schmitt, "Data-driven technologies and artificial intelligence in circular economy and waste management systems: a review," in *2021 IEEE International Symposium on Technology and Society (ISTAS)*, 2021, pp. 1–5.

[5] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[6] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," *Trends in Cognitive Sciences*, vol. 26, no. 2, pp. 174–187, 2022.

[7] Y. Lu, H. Wang, and W. Wei, "Machine learning for synthetic data generation: A review," 2023.

[8] S. I. Nikolenko, *Synthetic data for deep learning*. Springer, 2021, vol. 174.

[9] H. K. Ekbatani, O. Pujol, and S. Segui, "Synthetic data generation for deep learning in counting pedestrians." in *ICPRAM*, 2017, pp. 318–323.

[10] H. S. Behl, A. G. Baydin, R. Gal, P. H. Torr, and V. Vineet, "Autosimulate:(quickly) learning synthetic data generation," in *Computer Vision–ECCV 2020: 16th European Conference Proceedings, Part XXII 16*, 2020, pp. 255–271.

[11] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," 2017.

[12] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," 2016.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[14] A. Dutta, A. Gupta, and A. Zissermann, "VGG image annotator (VIA)," http://www.robots.ox.ac.uk/~vgg/software/via/, 2016.

[15] Z. Liu, B. Xu, B. Cheng, and X. Hu, "Interpolation parameters in inverse distance-weighted interpolation algorithm on dem interpolation error," *Journal of Sensors*, vol. 2021, pp. 1–14, 12 2021.

[16] F. Massa and R. Girshick, "maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch," https://github.com/facebookresearch/maskrcnn-benchmark, 2018.