Advanced Single-View Image-Based Framework for Volume Estimation in Urban Solid Waste Management

Julio Leite Azancort Neto[1], Romário da Costa Silva[1], Thalita Ayass de Souza[1], Carlos André de Mattos Teixeira[1], Evelin Helena Silva Cardoso[1], Jasmine Priscyla Leite de Araújo[1], Carlos Renato Lisboa Francês[1]

Instituto de Tecnologia – Universidade Federal do Pará (UFPA)[1]
Belém - PA - Brazil

{julio.azancort.neto, carlos.mattos}@itec.ufpa.br,
romario.silva@castanhal.ufpa.br, thalita_ayass@hotmail.com,
ehs.cardoso@gmail.com,{jasmine, rfrances}@ufpa.br

Abstract. *Efficient solid waste management is crucial for making the city a clean and sustainable environment. This paper presents a methodology composed of well-established algorithms for volume estimation in urban solid waste management using single-view images. The proposed system was built using state-of-the-art model-based algorithms including instance segmentation, depth estimation, and point cloud-based volume calculation The methodology demonstrates the ability to accurately estimate the volume of individual and multiple plastic bags containing municipal solid waste. We evaluated our approach using real-world data. Numerical results showed that the proposed system is promising even in complex scenarios. Despite challenges, such as: manual distance rescaling and limited datasets, our system holds considerable potential for further refinement and enhancement targeting scenarios as complex as real urban environments. The proposed methodology contributes to advancing management technologies in smart cities.*

Resumo. *A gestão eficiente de resíduos sólidos é crucial para tornar a cidade um ambiente limpo e sustentável. Este artigo apresenta uma metodologia composta por algoritmos bem estabelecidos para a estimativa de volume na gestão de resíduos sólidos urbanos usando imagens de visualização única. O sistema proposto foi feito a partir de algoritmos baseados em modelos de última geração, incluindo segmentação de instâncias, estimativa de profundidade e cálculo de volume baseado em nuvem de pontos. A metodologia demonstra a capacidade de estimar com precisão o volume de sacolas plásticas individuais e múltiplas, contendo resíduos sólidos urbanos. Avaliamos nossa abordagem utilizando dados do mundo real. Resultados numéricos mostraram que o sistema proposto é promissor mesmo em cenários complexos. Apesar dos desafios como reescalonamento manual de distância e conjuntos de dados limitados, nosso sistema possui um potencial considerável para refinamento e aprimoramento adicionais visando cenários tão complexos quanto cenários urbanos reais. A metodologia proposta contribui para o avanço das tecnologias de gestão em smart cities.*

1. Introduction

The generation of waste is a consequence of the increasing population, urbanization, and economic development. Despite the Solid Waste Management (SWM) problem, that

affects every individual and government in every country in the world. All this inadequately handled waste directly affects both public health and the environment [Arbeláez-Estrada et al., 2023].

Since the major waste disposal methods include collection, treatment, recycling and final disposal, hazards and risks from short-term contamination must be prevented [Azancort Neto et al., 2021]. Despite the National Solid Waste Policy demanding the end of dumps through the country of Brazil, in the state Pará, the landfill located in the municipality of Marituba receives solid waste from the capital Belém and its metropolitan region, that together collect around 40 thousand tons per day [Brito et al., 2020].

There is no doubt about the importance and challenge in the process of efficient waste management. The Integrated Sustainable Waste Management (ISWM) model shows all the necessary parts for an efficient and responsible process of waste disposal. The model is based on five points: Collection, transfer and transport; Generation and separation; Treatment; Recycling and Final disposal [Guerrero et al., 2013].

However, there are three very important items to be analyzed before the collection process, which is not commented on by the ISWM model. The process of identifying the location, volume and type of waste to be collected. Furthermore, accurately estimating the volume of waste is essential for efficient collection, transportation and processing. Technology can become an important instrument of environmental education, capable of minimizing environmental problems as it allows communication between responsible agents [Oliveira et al., 2019].

This paper presents a methodology for validating the estimation of the volume of solid waste present in garbage bags using only single-view images. Our system leverages cutting-edge model-based algorithms, including instance segmentation, depth estimation and point cloud-based volume calculation. This robust approach demonstrates the capability to accurately estimate the volume of both individual and multiple solid waste objects within an image. The implementation of this system can directly impact various aspects, including social-environmental factors, health, ecology, and the economy of the region where it is applied, contributing to a smarter and more sustainable city.

2. Related Work

For years now, different authors have used distinct approaches in identifying different types and pin-pointing the location of solid waste, all in an automatic manner. The most commonly used system is Machine Learning (ML), especially with Convolutional Neural Network (CNN), as used by Mao et al. (2022). Other techniques like infrared (IR) cameras by (Calvini et al. 2018) and support vector machines (SVM) by (Korucu et al. 2016), are also used.

Ozdemir et al. (2021) analyzed ML algorithms used in recycling systems. The reviewed techniques were: CNN, SVM, K-nearest neighbor and Artificial Neural Network (ANN). According to the authors, the black-box nature of the models and a large amount of meaningful data required for training are some of the biggest obstacles presented in this field.

Representing the CV research field, we have Lu and Chen (2022). They critically reviewed the sorting of municipal solid waste (MSW) using computer vision-based methods. In their paper, the author encountered limitations: simplified datasets lacked

real-world complexity and public datasets were scarce. Furthermore, the visible-image based approach struggled to distinguish materials with similar appearances.

Nevertheless, waste volume estimation has not been a very popular research theme. This may be due to the fact that the state of the art of volume estimation is not yet completely defined. Although, different authors have been using completely diverse methodologies, some based on Computer Vision (CV) and others in ML based-models. The background complexity and lack of quality and representative data are problems that affect every type of volume estimation method, solid waste is no different.

In summary, model-based, volume estimation methods often face limitations due to the challenge in obtaining high-quality ground truth data. Image depth estimation, for example, requires high-fidelity image capture. These limitations prevent existing methods obtaining accurate and consistent quantitative estimates [Alexandros Graikos et al. 2020]. To overcome these challenges, this work presents a hybrid deep learning algorithm that exploits the strengths of different methods to obtain the best volume object estimation. This framework aims to overcome the limitations of previous methods in order to improve performance in terms of accuracy and efficiency.

3. Methodology

The methodology outlined in this paper, aims to enhance the precision of volume estimation from single-view object images through the utilization of cutting-edge model-based algorithms. The proposed system pipeline can be divided into three distinct parts: segmentation network, depth estimation and point cloud-based volume estimation, Figure 1.

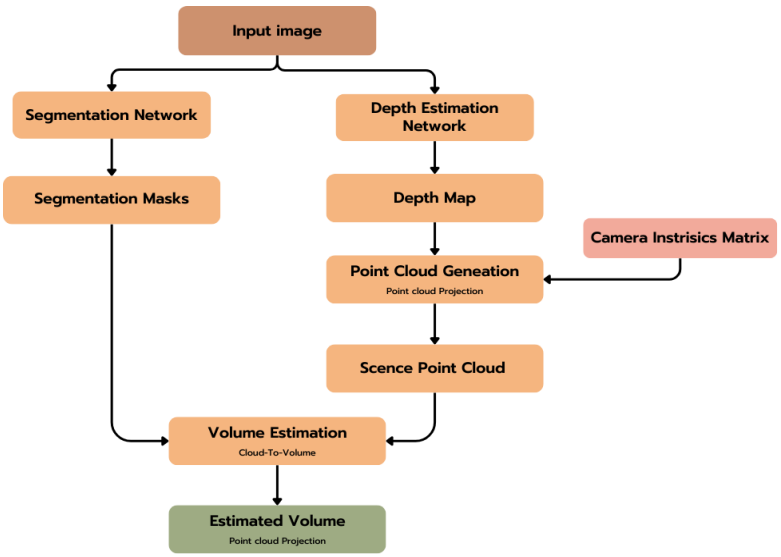

Figure 1. Proposed Framework adapted from Alexandros Graikos et al. (2020)

3.1. Waste Segmentation Network

The model chosen for the instance segmentation network was the Mask R-CNN [He et al. 2017], an extension of the Fast Region-based CNN [Girshick 2015], used for object detection. With this approach, we are able to locate multiple solid waste objects present in an image and predict an individual segmentation mask for each instance, so we can estimate each instance volume separately. Although not used in this work, the model is also able to discern between the different solid waste types present in a given image.

In line with Alexandros Graikos et al. (2020), who employed the COCO dataset [Lin et al. 2014] weights for their segmentation network, we adopted a similar approach by initializing our model with these weights. However, our refinement process involved fine-tuning the network using a distinct dataset composed of solid waste objects that we curated, since we were not able to find a usable and ready to use dataset for our application.

The curated dataset is composed of 278 images of solid waste in plastic bags, since this is the most common form of garbage dumping here in Brazil. The dataset is very simple, only with the segmentation masks for the analyzed objects and since we are only using one type of solid waste, we ignore and don't use any objects labels or types, except a generic name applied to all called "plastic", since we used plastic bags as our only analyzed object.

(a) (b)

Figure 2. Single (a) and multiple (b) solid waste examples contained in the gathered dataset.

3.2. Depth Estimation Network

This paper adopts the network architecture introduced by Godard et al. (2018) for the depth estimation task. In their study, a depth estimation network is trained exclusively on monocular video sequences, where each step involves the utilization of three consecutive frames (I_{T-1}, I_T, I_{T+1}) from the video for training purposes. The depth prediction network generates a depth map (D_T) for the input frame (I_T). Simultaneously, a pose estimation network produces the camera pose transformations ($T_{t\to t-1}, T_{t\to t+1}$) representing the relationships between the current frame and its adjacent frames. Utilizing the predicted depth map, pose transformations, and the known camera intrinsic matrix (K), the synthesis of the center frame occurs by sampling from the previous and next frames.

$$I_{t-1\to t} = I_{t-1}\langle proj(D_t, T_{t\to t-1}, K)\rangle \quad (1)$$
$$I_{t+1\to t} = I_{t+1}\langle proj(D_t, T_{t\to t+1}, K)\rangle$$

Here, "proj" refers to the coordinate projection method detailed by Zhou et al. (2017), and "⟨ ⟩" denotes the sampling operator. The ultimate training loss comprises the sum of a photometric loss (L_P) measuring the disparity between the synthesized and original images, along with a depth smoothness error (L_S). This depth smoothness error is expressed as a function that evaluates the smoothness of the predicted depth map.

$$L = L_P + \lambda L_s, \quad (2)$$

This approach encounters difficulties in producing a meaningful training signal when the actual pose transformations are zero. In such instances, the predicted depth values have no impact on the image synthesis process. This constraint narrows down the selection of videos suitable for training the network to those featuring substantial motion between frames. This limitation can be a difficult one to overcome, since solid waste videos may exhibit limited inherent motion, depending on how the capture is being made, making it challenging to ensure a robust training signal for the network, as the absence of significant motion between frames hampers the effectiveness of training mechanisms.

This limitation precluded us from training our own model. As previously mentioned, suitable datasets for this task are scarce. Gathering such data is not only difficult and time-consuming, but it can also pose health risks if handled by unqualified personnel.

Since Alexandros Graikos et al. (2020) also adopts the use of the network architecture introduced by Godard et al. (2018), we used his model and weights used for depth estimation. Although in their article, the author used the model specifically used for food volume estimation, in our tests, the model trained using the EPIC-KITCHENS dataset [Damen et al. 2018], that comprises over fifty hours of egocentric videos capturing food-handling activities and later fine-tuned by the authors using 38 videos capture by commercial smartphone cameras, demonstrated promising results by comparing volume estimation outcomes with the known volumes of solid waste objects.

3.3. Volume Estimation

Following the exploration of various volume estimation models, algorithms, and techniques in previous works, this paper adopts the method presented by Alexandros Graikos et al. (2020). As proposed by the authors, we leverage the depth map (D) extracted from the input image and the camera's intrinsic matrix (K) to project each pixel (x, y) onto its corresponding 3D point in space. This projection is achieved using homogeneous coordinates and the inverse projection model, resulting in a point cloud representation (P).

$$P_{xy} = K^{-1} \, [x \; y \; 1]^T \, D_{xy}, \tag{3}$$

Therefore, to enhance the differentiation among various solid waste objects and partition the set (P) into distinct subsets of solid waste points, we employed the segmentation mask generated by the instance segmentation network. From that, preprocessing of each point set commences with outlier removal utilizing a statistical outlier removal (SOR) filter. Subsequently, principal component analysis (PCA) is employed to identify the primary plane upon which the analyzed object resides. Following the PCA, the eigenvector corresponding to the minimal eigenvalue is selected to represent the normal vector of the base plane on which the object rests. To ensure consistency in object orientation, an additional step is implemented to guarantee that the plane is positioned at the object's bottom.

Although Alexandros Graikos et al. (2020) that proposed this methodology uses a planar plate for the volume estimation, we completely ignore this part of the suggested approach and use the base, usually the ground, as our plane base. Knowing that this can be improved but is not as simple as suggested in the original paper, since there is no default behavior in incorrect waste disposal. Subsequently, the projected points are used

to construct an α-complex derived from the Delaunay triangulation [Edelsbrunner & Harer 2010]. This α-complex partitions the covered area on the base plane into a collection of triangles. Then, the estimated volume is defined by the average of each triangle vertex from the analyzed solid waste object.

Since the videos used for the training of the depth network lack the ground truth depth information, the depth predictions are not in a metric scale. To address the absence of ground truth depth in the training process, we adopt the median ground truth rescaling technique from Zhou et al. (2017). This method scales the predicted depth map (D) by a constant factor.

$$s = \frac{median(D^{gt})}{median(D)}, \tag{4}$$

In our experiments, the scaling factor is determined by the median ground truth depth (median(D^{gt})) which approximates the distance between the camera sensor and the food object. For our case, we assumed it to be 0.5 meters for all test cases. However, this value can be manually and easily modified, depending on the applied scenario and distance to the analyzed object.

As commented before, the depth predictions are not in a metric scale, so the formula of median ground truth rescaling technique proposed by Zhou et al. (2017) and described in (3), is used to obtain the depth value. For the training of the segmentation network using the Mask R-CNN algorithm. The input batch size, learning-rate and the same data augmentations applied in the depth estimation network. The parameters used can be found in Table 1.

Table 1. Parameters utilized in the Segmentation and Depth Estimation Network.

Algorithm	Parameter	Value
Segmentation	Batch size	1
	Detection min confidence	0.7
	Detection NMS threshold	0.3
	Learning rate	0.001
	Validation steps	51
	Backbone	resnet101
	Training epochs	2
Depth Estimation	Input resolution	128x224
	Depth outputs range	0.01 to 10
	Smoothness Term	10^{-2}
	Training Epochs	20
	Learning Rate	10^{-4}
	Ground Truth Expected Median Depth	0.50

The model was then trained with the pre-trained weights of COCO dataset [Lin et al. 2014] on a split of 221 used for training, 28 for validation and 29 for testing the volume

estimation. All items had the same class called “plastic”, since the main object was plastic bags, which is the most common way to throw household solid waste. In the volume inference, we change the value of the field-of-view angle of the camera sensor, which was changed from 70° to 79.5°, this value is to generate the same intrinsic data. We did not change the values of the Z-Score for the SOR filter or α-complex.

5. Results, Difficulties and Discussions

To assess the proposed system, we measured the volume of 29 plastic bags that we are going to name “Domestic waste”. Each domestic waste had 2 or 3 images taken, either on a different angle or in a different distance from the analyzed object. For that matter, each type of domestic waste may have multiple Relative Percentage Error (RPE) analyzed.

For the real measurement we calculated all plastic bags as an irregular polygon, more specifically a rectangle. Although some works use methods like water displacement, this approach not only fails to accurately represent the true value of the object under analysis but also isn't suitable for the type of objects we are handling. To evaluate the real results of the volume estimation of solid waste, in Table 2, we present the RPE of each image of the analyzed type, and the respective mean absolute percentage errors (MAPE) of the images with a single instance of solid waste, and in Table 2, with multiple instances.

Table 2. Real volume measured, the RPE of each variation and the MAPE estimated from single and multiple instances of solid waste.

Type	Volume (L)	RPE 1 (%)	RPE 2 (%)	RPE 3 (%)	MAPE (%)
Single Domestic Waste 1	0.9285	0.49	3.18	2.29	1.98
Single Domestic Waste 2	0.9936	10.29	1.05	39.19	16.84
Single Domestic Waste 3	1.6128	58.94	5.54	8.11	24.19
Single Domestic Waste 4	0.5440	36.98	13.18	11.88	20.68
Multi. Domestic Waste 1	2.2495	2.65	5.33	33.33	13.77
Multi. Domestic Waste 2	3.146	1.71	4.51	19.87	8.69
Multi. Domestic Waste 3	3.6752	5.94	10.10	26.25	14.09
Multi. Domestic Waste 4	3.4532	3.99	9.63	14.63	9.41

When the segmentation network divided a single object into its component parts, we addressed this by summing the predicted volumes of each individual segment. Since the proposed system is capable of identifying multiple volumes, we tested this feature by estimating the volume of multiple plastic bags in one image.

In cases where the bags overlapped, we had two different but expected results. In one of the cases, the algorithm tried to estimate the volume for every single instance of the solid waste object. In other cases, the system used all solid waste instances together, calculating them all as a single value, this method had better results in multi-instance volume estimation. More useful data might enhance the decision-making process when employing these two types of analysis.

Given the lack of a generalist solution to the object volume estimation problem, we evaluated our framework's performance by comparing its results to those presented in

state of the art. In a head-to-head comparison, both our single-instance and multiple-instance MAPE scores outperformed those achieved by the authors. Notably, we achieved this with a less complex algorithm, as our framework relies solely on the algorithms described earlier and avoids the use of external aids employed by the authors. Despite having a smaller test set, our single-instance MAPE yielded a mean MAPE of 15.92%, compared to 18.72% for the best four results. Similarly, our multiple-instance mean MAPE achieved 11.49%, significantly lower than the 32.07% reported by Alexandros Graikos et al. (2020), for example. Our results also showed a max overall MAPE value of 24.19% and min of 1.98%, while the referenced paper results had a max of 108.30% and min of 15.85%. Therefore, our experiments revealed that when objects overlap, segmenting all instances together as a single large item yielded superior results. While this approach might hinder precise individual volume determination, the overall accuracy gains outweigh this limitation.

One of the primary shortcomings of this approach, as well as others that do not utilize sensors like LiDAR [Li et al. 2022], is the dependency of manually applying the correct median distance rescaling. The wrong applied value can lead to wrongly volume estimation and totally ruins any automations methods. By extracting the median depth in an automated manner, we are able to change and adapt the median depth value on the go. Sensors like LiDAR are often used for this purpose, especially in industry.

The biggest challenge lies in acquiring pertinent data. Despite our efforts, we were unable to locate a dataset that met our requirements and was readily available. Consequently, we had to undertake the time-consuming task of creating and segmenting our own data. While this process is slow, it is indispensable. It underscores the critical need for additional data, particularly for training the depth estimation network. As mentioned previously, dealing with images containing multiple instances of the analyzed object that overlap presents a significant challenge. In such scenarios, the algorithm struggles to accurately estimate the correct value, further complicating the analysis process.

Overall, the proposed system showed promising results, especially considering the potential for further improvements through training our own models and adjusting weights beyond the segmentation aspect. The results align well with the approximate error values reported in the literature using similar models. Finally, with the capability to estimate volumes through single-view images, the model can be applied to the development of low-cost real-world applications, such as mobile apps or embedded systems in waste collection services.

6. Conclusion

In this work, we presented a methodology that aims to validate the process of volume estimation applied in the realm of solid waste management using only single-view images. The proposed system used state of the art model-based algorithms such as instance segmentation, depth estimation and point cloud-based volume estimation. This robust system was capable of accurately estimate the volume of plastic bags containing municipal solid waste.

In our results, we showcased the effectiveness of the proposed approach. We were able to achieve promising results even in scenarios involving multiple instances of overlapping objects. Although we encountered some challenges such as the manual

application of correct median distance rescaling and the scarcity of readily available datasets, our system demonstrates considerable potential for further refinement and enhancement. We recognize the importance of meaningful and high-quality data. With more data we will be able to train and optimize our model and weights to better suit the specific needs of our application and the overall system effectiveness and results.

Despite the framework demonstrating effectiveness and promising results, it requires further improvements for real-world applications. A key area for improvement, beyond the need for high-quality data, is automating the median distance rescaling process. This automation would allow the framework to analyze objects at varying distances, leading to better overall generalization of the methodology. Consequently, this would improve the Root Mean Squared Prediction Error (RPE) and the Mean Absolute Percentage Error (MAPE) results.

In general, the proposed system has demonstrated significant advancements in waste management technology, providing solutions for accurate quantitative estimates with environmental protection, aligning with the evolving needs and requirements for sustainable waste disposal practices. The continuation of this article aims to the automation process of rescheduling the average distance using model-based sensors or technologies such as LiDAR, for instance. Furthermore, it is expected that with this improvement, the problem of generalization present in current volume estimation methodologies can be overcome. By addressing this issue, we anticipate that the framework will become applicable within the context of smart cities.

References

Alexandros Graikos, Charisis, V. S., Dimitrios Iakovakis, Stelios Hadjidimitriou, & Hadjileontiadis, L. J. (2020). Single Image-Based Food Volume Estimation Using Monocular Depth-Prediction Networks. Lecture Notes in Computer Science, 532–543. https://doi.org/10.1007/978-3-030-49108-6_38.

Arbeláez-Estrada, J. C., Vallejo, P., Aguilar, J., Tabares-Betancur, M. S., Ríos-Zapata, D., Ruiz-Arenas, S., & Rendón-Vélez, E. (2023). A Systematic Literature Review of Waste Identification in Automatic Separation Systems. Recycling, 8(6), 86. https://doi.org/10.3390/recycling8060086.

Azancort Neto, J. L., Gonçalves, A. L. S. ., Cruz, B. C. C. da ., Gomes, L. L. ., & Costa , D. C. L. . (2021). Artificial Intelligence implemented to recognize patterns of sustainable areas by evaluating the database of socioenvironmental safety restrictions. Research, Society and Development, *10*(10), e212101018841. https://doi.org/10.33448/rsd-v10i10.18841.

Brito, D. A. C., Seabra, L. C., Lima, P. D. M., & Souza, C. M. N. (2020). Manejo de Resíduos Sólidos e de Águas Pluviais: O (Des)Controle Social em Belém, Pará. Revista Eletrônica de Gestão E Tecnologias Ambientais, 8(2), 103. https://doi.org/10.9771/gesta.v8i2.42221.

Calvini, R., Orlandi, G., Foca, G., & Ulrici, A. (2018). Development of a classification algorithm for efficient handling of multiple classes in sorting systems based on hyperspectral imaging. Journal of Spectral Imaging. https://doi.org/10.1255/jsi.2018.a13.

Damen, D., Doughty, H., Giovanni Maria Farinella, Fidler, S., Furnari, A., Kazakos, E., Davide Moltisanti, Munro, J., Perrett, T., Price, W., & Wray, M. (2018). Scaling Egocentric Vision: The "Equation missing" Dataset. 753–771. https://doi.org/10.1007/978-3-030-01225-0_44.

Edelsbrunner, H., & Harer, J. (2010). Computational Topology: An Introduction. American Mathematical Society, Providence.

Girshick, R. (2015). Fast R-CNN. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1504.08083.

Godard, C., Oisin Mac Aodha, Firman, M., & Brostow, G. J. (2018). Digging Into Self-Supervised Monocular Depth Estimation. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1806.01260.3

Guerrero, L. A., Maas, G., & Hogland, W. (2013). Solid waste management challenges for cities in developing countries. Waste Management, 33(1), 220–232. https://doi.org/10.1016/j.wasman.2012.09.008.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. ArXiv.org. https://arxiv.org/abs/1703.06870.

Korucu, M. K., Kaplan, Ö., Büyük, O., & Güllü, M. K. (2016). An investigation of the usability of sound recognition for source separation of packaging wastes in reverse vending machines. Waste Management, 56, 46–52. https://doi.org/10.1016/j.wasman.2016.06.030.

Li, N., Ho, C. P., Xue, J., Lim, L. W., Chen, G., Fu, Y. H., & Lee, L. Y. T. (2022). A Progress Review on Solid-State LiDAR and Nanophotonics-Based LiDAR Sensors. Laser & Photonics Reviews, 16(11), 2100511. https://doi.org/10.1002/lpor.202100511.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014, 8693, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Lu, W., & Chen, J. (2022). Computer vision for solid waste sorting: A critical review of academic research. Waste Management, 142, 29–43. https://doi.org/10.1016/j.wasman.2022.02.009.

Mao, W.-L., Chen, W.-C., Fathurrahman, H. I. K., & Lin, Y.-H. (2022). Deep learning networks for real-time regional domestic waste detection. Journal of Cleaner Production, 344, 131096. https://doi.org/10.1016/j.jclepro.2022.131096.

Oliveira, M., Silva, J., Silva, R., & Teran, L. (2019). Aplicação web para Gerenciamento de Resíduos Sólidos Recicláveis. In Anais do X Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais, (pp. 145-153). Porto Alegre: SBC. doi:10.5753/wcama.2019.6429

Zhou, T., Brown, M. A., Snavely, N., & Lowe, D. (2017). Unsupervised Learning of Depth and Ego-Motion from Video. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1704.07813.