Using Synthetic Images to Detect Groceries on Shelves

Andrea Bragagnolo[1,*], Gianluca Dalmasso[1] and Andrea basso[1]

[1]*Synesthesia s.r.l, Turin, Italy*

Abstract

Object detection on grocery store shelves has become essential in retail industries to ensure planogram compliance: arranging products in a specific layout to maximize sales and improve customer experience. Neural networks have been widely employed for this task, achieving excellent performances. However, standard datasets may not be adequate for this task as they are often region-specific or too generic, making it difficult to distinguish between different products. To address this issue, we propose using synthetic images to create a more diverse and comprehensive dataset. Our results show that synthetic images can successfully train a detection system resilient to challenging scenarios (e.g., occlusions and variations in lighting conditions).

Keywords

Retail, Deep Learning, Synthetic Data

1. Introduction

Automation, especially deep learning, in the retail context, has become an important area of research in recent years [1, 2], with applications ranging from ensuring products are arranged on shelves correctly and keeping track of how much inventory there is. Neural networks have been used to detect objects on shelves [3, 4] but require large amounts of labeled data to be trained effectively. While datasets such as GroZi-120 [5], SKU-110k [6], and many more have been proposed for this purpose, they are comprised of region-specific products or the labeling is too generic (i.e., every object belongs to the same class), making it difficult to distinguish between different local products accurately.

Synthetic data is a popular solution for addressing the challenge of limited hand-labeled data required to train intelligent systems. By generating synthetic data, diverse and comprehensive datasets can be set up, simulating various scenarios and conditions that may be difficult to capture in real-world datasets. This can ultimately improve the performance and generalizability of the trained models. For instance, in autonomous driving, synthetic datasets can be effectively used to train neural networks with high accuracy [7, 8, 9].

This paper presents our approach for generating synthetic images to train a YOLOv5 [10] neural network for object detection on shelves. We demonstrate the effectiveness of our approach through experimental results, and we show that our system achieves high accuracy in detecting different products on shelves, demonstrating that our model, trained on synthetic datasets, can perform well even on real images. Overall, our approach provides a valuable contribution to the field of object detection in the retail context. It offers a practical solution for addressing the challenges posed by limited datasets.

2. Facing the lack of labeled data

To train our neural network to detect groceries on shelves, we generated synthetic images of retail store shelves using the BlenderProc library [11]: a Python library that enables the automation of the rendering process in Blender, a popular open-source 3D modeling software. This tool provides a flexible and efficient framework for generating large-scale datasets of synthetic images with diverse configurations and conditions.

2.1. Image synthesis

To create a diverse and comprehensive dataset, first, we faithfully modeled each product that the network should detect; then, for each render, we randomized various elements of the scene, including the shelf, the products, their position and rotation, the camera point of view, and the illumination. By randomizing these elements, we could generate many images with different configurations and conditions to simulate real-world scenarios. To increase the specificity of the neural network and reduce the network's false positive detections, we also added negative samples, i.e., products that should not be recognized, such as objects with features similar to our targets or unrelated items.

2.2. Automated labeling

The annotation of products in the renders is done automatically by BlenderProc using a built-in algorithm that uses the information of the 3D model to project the

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

[*] Corresponding author.

andrea.bragagnolo@synesthesia.it (A. Bragagnolo); gianluca.dalmasso@synesthesia.it (G. Dalmasso); andrea.basso@synesthesia.it (A. basso)

bounding boxes into the 2D image plane. Of all the generated bounding boxes, we keep only those corresponding to products with at least 50% of the area visible in the render. Besides the box information, each label includes the product's class (i.e., the EAN). This automated annotation process saves time and resources, eliminating the need for manual annotation. Additionally, it ensures consistency and accuracy in the annotation process, as it is performed uniformly for all images in the dataset.

Figure 1 shows some examples of the synthetic images we generated using our system. The rendered images display a retail store shelf with various products the model aims to recognize. The positive examples, i.e., the products the network should detect, are annotated with bounding boxes, which vary in color based on the object's class. The images also include negative examples, although without ground truth information. The examples are presented in different orientations, positions, and configurations, and the scene is superimposed on a randomly selected background image.

2.3. Neural network training

We used images at standard COCO resolution to train our YOLO network. However, to generate the synthetic images for training, we create them with a higher resolution. The reason for this is that the higher resolution allows us to capture more details and information about the scene, which can improve the performance and accuracy of the trained model. Additionally, by rendering at a higher resolution, we can downsample the images to the desired training resolution without losing too much information or detail. This approach ensures the training images retain sufficient quality and clarity, even after downsampling.

To increase the robustness of our trained YOLO model and improve its generalization capabilities, during training time, we added a random background to the shelf renders. By doing so, we increased the variance of the training data and reduced bias toward specific background patterns or textures. This ensures that the model is invariant to the background around the shelf when applied to natural images, as it has learned to detect the products regardless of the background. To achieve this, we used a collection of high-resolution images of various backgrounds, and we randomly selected and inserted one of these images behind the shelf in each render. Adding this extra variability to the training data has improved the model's ability to detect products in a wide range of real-world environments.

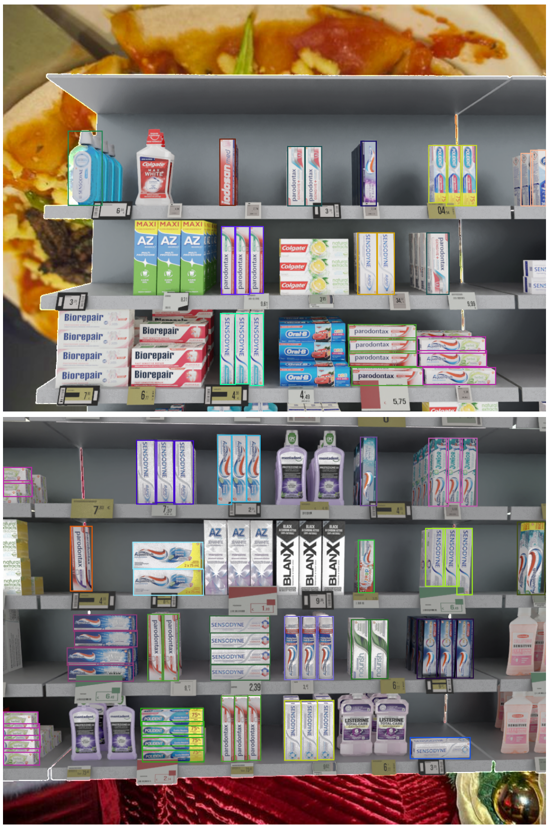

Figure 1: Example of shelf renders. Different products (both positive and negative samples) are displayed in various positions. Ground truth bounding boxes are shown in different colors according to the object's class. Random backgrounds are applied to the renders at training time.

3. Result of synthetic training on real samples

To test the model trained using synthetic images, we used a hand-labeled dataset with a total of 73 different classes. During our testing, we focused on dental hygiene products. The results of this study demonstrate that the object detection model can generalize well to natural images. The model achieved good performance on real images with an average Precision of 0.846 and an average Recall of 0.883, indicating that the model can accurately identify objects in various settings.

Figure 2 shows detection results on a sample of natural images. The model can detect objects accurately in various settings and lighting conditions. Overall, these results suggest that utilizing synthetic data to train object detection models has the potential to be a powerful tool for improving the accuracy and efficiency of computer vision systems.

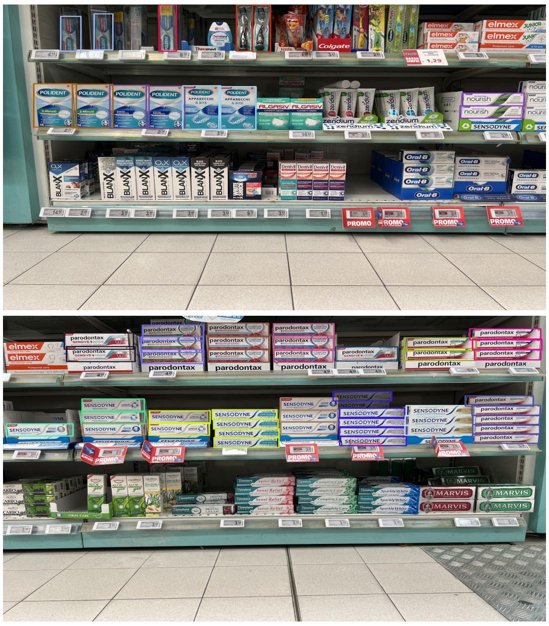

Figure 2: Example of detection on real images. Predicted bounding boxes are shown in different colors according to the predicted class.

4. Conclusions

We presented an approach for generating synthetic images to train a YOLOv5 neural network for object detection of groceries products. Our approach relies on using BlenderProc to automate the rendering and labeling process in Blender, creating a diverse and comprehensive dataset of synthetic images that can simulate various scenarios and conditions. Our system achieved high accuracy in detecting different products on shelves, and our experiments demonstrated that the model trained on synthetic datasets could perform well on real images. Our approach provides a valuable contribution to object detection in the retail context, offering a practical solution for addressing the challenges posed by limited datasets. Future work may focus on refining the generation process and expanding the model's capability to detect additional products or improve its speed and efficiency.

5. Acknowledgement

We thank the Computer Science Department of the University of Turin for its support throughout the duration of this project both in expertise and equipment that allowed us to conduct our experiments.

References

[1] G. Varol, R. S. Kuzu, Toward retail product recognition on grocery shelves, in: Sixth International Conference on Graphic and Image Processing (ICGIP 2014), volume 9443, SPIE, 2015, pp. 46–52.

[2] D. Grewal, A. L. Roggeveen, J. Nordfält, The future of retailing, Journal of Retailing 93 (2017) 1–6. URL: https://www.sciencedirect.com/science/article/pii/S0022435916300872. doi:https://doi.org/10.1016/j.jretai.2016.12.008, the Future of Retailing.

[3] A. De Biasio, Retail shelf analytics through image processing and deep learning, Universityo Padua, Padua, Italy (2019).

[4] A. Tonioni, E. Serra, L. Di Stefano, A deep learning pipeline for product recognition on store shelves, in: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), IEEE, 2018, pp. 25–31.

[5] M. Merler, C. Galleguillos, S. Belongie, Recognizing groceries in situ using in vitro training data, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8. doi:10.1109/CVPR.2007.383486.

[6] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, T. Hassner, Precise detection in densely packed scenes, in: Proc. Conf. Comput. Vision Pattern Recognition (CVPR), 2019.

[7] S. R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, 2016, pp. 102–118.

[8] A. Gaidon, Q. Wang, Y. Cabon, E. Vig, Virtual worlds as proxy for multi-object tracking analysis, in: CVPR, 2016.

[9] S. R. Richter, Z. Hayder, V. Koltun, Playing for benchmarks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2213–2222.

[10] G. Jocher, YOLOv5 by Ultralytics, 2020. URL: https://github.com/ultralytics/yolov5. doi:10.5281/zenodo.3908559.

[11] M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. Katam, A. Lodhi, A. Penzkofer, BlenderProc2, 2021. URL: https://github.com/DLR-RM/BlenderProc/.