Automatic Image Dataset Generation for Footwear Detection

Javier Martínez-Cesteros, Gonzalo López-Nicolás

Grupo de robótica, percepción y tiempo real (RoPeRT)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: *598118@unizar.es*

Abstract

In this work, we explore the problem of object detection in the shoe manufacturing application. We propose different methods for the automatic or semi-automatic generation of image datasets. Our approach avoids the time consuming task of collecting labelled images, which is required for training neural networks.

Introduction

The design of automatic algorithms for object detection in images is a very difficult task. However, object detection is a fundamental ability for improving human work conditions by automating harmful, heavy and repetitive industrial tasks. Object detection is even more difficult with deformable objects or when there is a great variety of products, such as in the footwear industry. In this application, automatic shoe detection is of paramount importance to improve the benefits of automatization in the footwear manufacturing.

Dataset generation for learning

To solve the object detection problem, we have used Convolutional Neural Networks, due to their capabilities in the computer vision field, which have reached impressive results in image classification compared with previous techniques. In particular, we have chosen Mask-RCNN [1] because of its good precision. Besides, that implementation is open source and the code is available in the repository of Matterport [2].

The chosen network allows the object detection, classification, semantic segmentation and instance segmentation, but it needs to be previously trained with the object of interest. The training process cannot be performed in the absence of image datasets with labelled objects. Although there are many available datasets, many objects are not considered. In particular, labelled datasets of shoes in our target application are difficult to find. Moreover, Mask-RCNN training requires *instance segmentation* of the objects in the dataset, but most public datasets only provide the *bounding box* of the labelled objects. Therefore, it is necessary to generate specific new image datasets.

In the classic approach, the dataset of images is generated by collecting pictures of shoes and labelling them, one by one. This task can be performed manually using programs such as VGG Image Annotator (VIA) [2]. This manual method requires unfeasible amount of time, it is not scalable, and any class variation or object addition implies re-labelling the images.

Here, we have developed an automatic dataset generation method that only needs images of the objects of the class to be identified with distinctive background. This method also requires a set of diverse images as backgrounds without any of the objects we wish to detect. In particular, we have used the Microsoft COCO dataset [4]. A diagram to illustrate this method is depicted in Fig. 1.

Our method randomly overlay the background images with the objects. In particular the user chooses the maximum number of objects to include per image. Then, a random number between one and this maximum will be determined for each image. In the overlay process, a random scale is chosen in a range proportional to each object and image sizes. The position of the object in the image is also chosen randomly, allowing partial clipping of the object at the edge of the image. Note that occlusions between objects are also considered. The flexibility of the method lies in being able to use each background with different objects with different sizes and positions obtaining as many resulting images as desired. This flexibility is fundamental for the dataset effectiveness in training networks.

Additionally, we propose another method in which, given background images, we choose manually the type of object, its position, scale and rotation to overlay the image with it. This method gives us the possibility to generate more realistic images than with the random method although it requires some manual positioning effort.

Experimental Results

The training process has been performed with a computer with 16 GB of RAM, Processor Intel® Core™ i7-8700 at 3192 MHz and the Titan Xp graphic card under Ubuntu OS. The training of the Mask-RCNN takes around five hours, and we have obtained the promising results shown in Fig. 2. We can see that the net is able to detect shoes of different shapes, viewpoints and orientations even with occlusions, with a good success rate.

Conclusions and Future Work

We conclude that the classical manual labelling presents similar results to the proposed methods of manual object positioning and the automatic random object positioning. The main advantage of the presented methods is that the required time to generate a training dataset is dramatically reduced.

The clear scalability of the proposal allows increasing the number of object classes to detect by training networks, such as different kinds of footwears (boots, sneakers, shoes, sandals, etc.), tools widely used in the manufacture process (hammer, file, polisher, etc.), and a variety of products and components involved in the production (glue, soles, insoles, etc.).

ACKNOWLEDGMENTS

This work was supported by project COMMANDIA SOE2/P1/F0638 (Interreg Sudoe Programme, ERDF) and project PGC2018-098719-B-I00 (MCIU/AEI/FEDER, UE). The Titan Xp used for this research was donated by the NVIDIA Corporation.

REFERENCES

[1]. HE, K., GKIOXARI, G., DOLLÁR, P., and GIRSHICK, R. Mask R-CNN. 2018. Available from: arXiv:1703.06870 [cs.CV].

[2]. ABDULLA, W. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. 2017. https://github.com/matterport/Mask_RCNN.

[3]. DUTTA, A., and ZISSERMAN, A. The VGG Image Annotator (VIA). 2019. arXiv:1904.10699 [cs.CV]

[4]. LIN, T.-Y., et al. Microsoft COCO: Common Objects in Context. 2015. arXiv:1405.0312 [cs.CV]

Fig. 2. Different examples of the obtained results

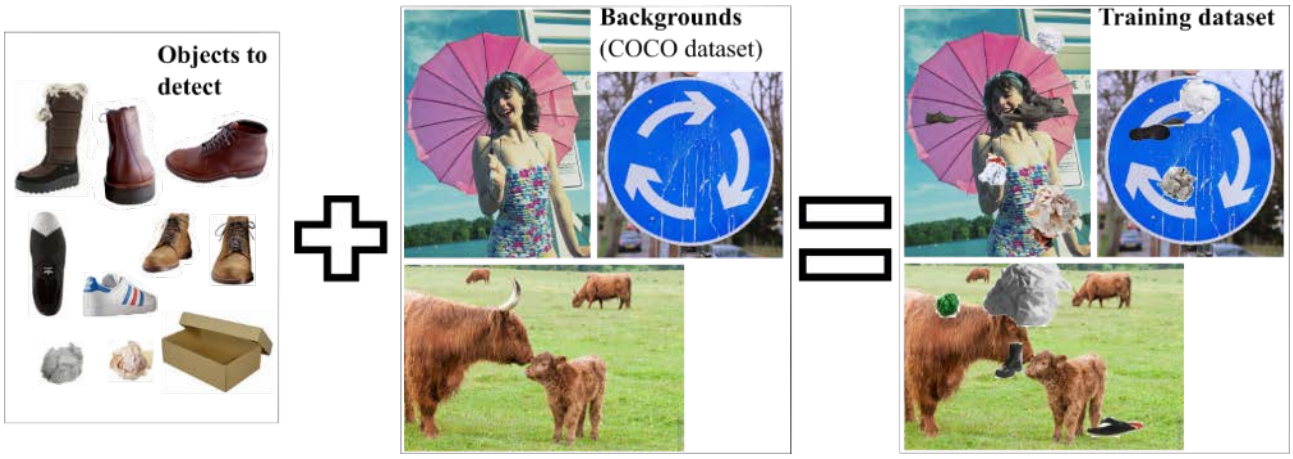

Fig. 1. Diagram of the first proposed method for automatic dataset generation. The dictionary of objects (left) is added to the images of the Microsoft COCO dataset (centre), obtaining our training dataset with the objects correctly labelled (right).