Robust Detection of Outdoor Urban Advertising Panels in Static Images

Ángel Morera[1], Ángel Sánchez[1(✉)], Ángel D. Sappa[2,3], and José F. Vélez[1]

[1] Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
{angel.morera,angel.sanchez,jose.velez}@urjc.es
[2] Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador
[3] Computer Vision Center, Bellaterra, Barcelona, Spain
sappa@ieee.org

Abstract. One interesting publicity application for Smart City environments is recognizing brand information contained in urban advertising panels. For such a purpose, a previous stage is to accurately detect and locate the position of these panels in images. This work presents an effective solution to this problem using a Single Shot Detector (SSD) based on a deep neural network architecture that minimizes the number of false detections under multiple variable conditions regarding the panels and the scene. Achieved experimental results using the Intersection over Union (IoU) accuracy metric make this proposal applicable in real complex urban images.

Keywords: Object detection · Urban ads panels · Deep learning · Single Shot Detector (SSD) architecture · Intersection over Union (IoU) metric · Augmented Reality

1 Introduction

The concept of *smart city* was coined twenty years ago [1] and it refers to a urban space that applies Information and Communication Technologies (ICT) to enhance the quality and performance of urban services such as energy, transportation and utilities in order to reduce resource consumption, wastage and overall costs. Urban space maintenance in smart cities is a challenging and time consuming task since this space is actually surrounded by or is embedded with smart IoT systems that can efficiently capture and process huge amount of data. Hence, in the smart city context, having an approach that allows a continuous monitoring of the urban space will help to envisage an efficient management of the public resources. The urban space monitoring problem has been recently tackled using frameworks based on citizens' sensing devices, under a crowdsourcing philosophy [10].

The outdoor advertising industry has experimented an important growth in recent years [2]. In streets of urban environments, ads panels and billboards are everywhere, and they are also the only media that drivers and pedestrians cannot

F. De la Prieta et al. (Eds.): PAAMS 2019 Workshops, CCIS 1047, pp. 246–256, 2019.
https://doi.org/10.1007/978-3-030-24299-2_21

escape (i.e., differently from other forms of publicity, outdoor advertising cannot be blocked by people). In consequence, this is one the most cost-effective forms of advertising available. Moreover, since current smartphones are equipped with a variety of embedded sensors like cameras, GPS or 3G/4G, it is possible to get closer the final user a variety of Augmented Reality (AR) applications [6]. This way, the citizens using their smartphones can better deepen and, perhaps, enjoy the contents associated to urban advertisements. Moreover, with the emergence of digital billboards/panels, the outdoor advertising industry is even more valuable since going digital gives advertisers the flexibility to schedule short and long-term publicity campaigns.

Advertising panels are a type of urban furniture that commonly presents a normalized shape and a more reduced size than billboards. Publicity panel detection in images presents important advantages both in the real world as well as in the virtual world. In the first case, after detection of panels, it is possible to recognize the product included in the publicity and get more information about it through AR applications. Moreover, it is possible to detect whether or not the advertised product information is currently updated. Regarding to the publicity in Internet urban images in applications like Google Street View, it would be possible, when detecting panels on these images, to replace the publicity that appears inside a panel by another one proposed by a financing company.

Automatic outdoor detection and localization in real urban outdoor images is a difficult task due to multiple variability conditions presenting images including those containing advertising elements like panels. For example, scene illumination conditions (solar daylight or artificial lights), panel perspective view, size ratio of present panels with respect to image size, complex scene background (i.e., presence of multiple elements surrounding the panels like buildings, vehicles and/or different infrastructures), among other factors. Figure 1 illustrates some of these involved difficulties in outdoor images containing advertising panels.

Visual detection and recognition problems, applied to specific elements, in outdoor images have been widely studied. For example, this is the case of vehicle localization [17], traffic sign detection [4] or car plates [11]. However, as far as we know, there are not published works on detecting outdoor ads panels. Related problems such as text and objects detection inside segmented billboard images [7] or the localization of billboards on streamed sport videos have been investigated [16]. Another related studied application is the insertion of virtual ads in street images based on localization of specific regions on them (e.g., buildings facades) [6].

Convolutional networks are deep artificial neural networks that are currently used in many Computer Vision tasks, such as classifying images into categories, detecting objects, clustering images by similarity or performing object recognition within scenes. The Single Shot Detector (SSD) [9] is based on a convolutional network that produces a fixed-size collection of bounding boxes and scores (related to the presence of object class elements in these boxes), followed by a non-maximum suppression stage to produce the final detections.

This paper proposes a robust method for the automatic detection of urban ads panels in outdoor images. The proposed solution is based on deep neural networks and it achieves a high accuracy in detection of panels under multiple and combined variabilites on illumination, position and size of detected targets. The number of false detections is reduced in order to allow a more practical application of the proposed approach.

The work is organized as follows. Section 2 describes a preprocessing stage applied to the given images, as well as the proposed deep architecture and its training. Section 3 offers some details on the dataset used and the experiments performed for detecting the panels in images. Finally, the last Section outlines the conclusions of this work.

2 Proposed Solution

This section summarizes the initial preprocessing applied to the images, the deep neural network used to detect outdoor panels and setup information on how this network was trained.

2.1 Image Preprocesing

Main preprocessing consists in rescaling the original images by preserving their aspect ratio. For such purpose the smaller side of an image was set to 512 pixels and the larger side was set to the proportional size in pixels, so that the aspect ratio is preserved. After that, the larger side is trimmed so that it would also be 512 pixels without losing any part of the panel. Therefore, all training, validation and test images were rescaled to 512×512.

2.2 Single Shot Detector (SSD) Architecture

Single Shot MultiBox Detector (SSD) networks were introduced by Liu and collaborators in 2016 [9]. A SSD network implements a method for detecting multiple object classes in images by generating confidence scores related to the presence of any object category in each default box. Moreover, it produces adjustments in boxes to better match the object shapes. This network model is suited for real-time applications since it does not resample features for bounding box hypotheses (like in models such as Faster R-CNN [13]). Additionally, SSD is as accurate as other single-shot approaches like YOLO [12]. The SSD architecture is based on a feed-forward convolutional network and its object detection approach has two steps: (1) extract feature maps, and (2) apply convolution filters to detect objects. SSD uses VGG16 [15] to extract feature maps. Then, it detects objects using the Conv4_3 layer. Each prediction is composed of a bounding box and 21 scores for each class (one extra class for no object); the class with highest score is selected as the one for the bounded object. Conv4_3 makes a total of $38 \times 38 \times 4$ predictions: four predictions per cell independently from depth of feature maps. Many predictions will contain no object as it is expected and uses the class '0' to indicate that none object was detected in the image. Figure 2 illustrates the typical layer structure of a SSD network.

Fig. 1. Some test images of panels including diverse variabilities: (a) night image; (b) panel rotated due to image perspective; (c) partial occlusion of panel; and (d) reduced size (and partial occlusion) of panel in the image.

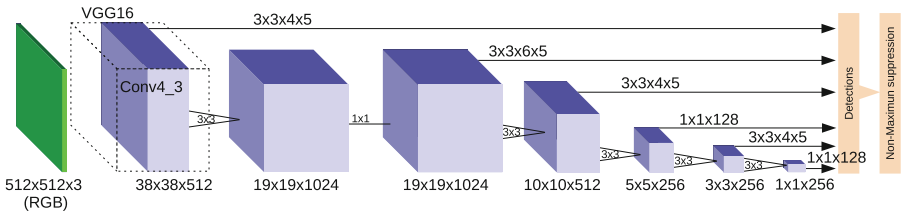

Fig. 2. Single Shot Detector (SSD) architecture model.

2.3 Training Details

SSD only needs an input image and the ground truth boxes for each object during training. In our approach, a SSD_MobileNet_v1, pretrained with Microsoft COCO dataset [8], was used. MobileNets [5] are a family of more efficient models including depth-wise separable convolutions, suitable for mobile and embedded vision applications.

The network input was adapted to the size of our preprocessed images. Then, it was finely tuned and trained using our own panel dataset (some details on the dataset are given in next section). In our problem, only one class was required (i.e. the 'panel' class) and the network itself can discriminate in images between what is a 'panel' and what is not. Experiments were performed on an standard computer and using an small batch size of 6–8 and different number of epochs so far of 180,000. Different values of learning rates varying from 0.004 to 0.001 were evaluated, with a momentum of 0.9. Approximately, a number of 5,800 panel images were used for training and 100 images for validation.

3 Results

This section describes the dataset used and the achieved experimental results. The influence of panels' sizes in images with respect to the accuracy on detections is also studied.

3.1 Dataset

As far as we know, there are not public datasets of outdoor urban panel images. In consequence, one of the first tasks in this project consisted in creating a dataset to train the SSD network. This dataset will be released to other researchers interested in the considered problem. We have firstly collected approximately 1,800 images which were separated into training and validation sets. Test images were collected separately, a number of 140 test images in total. Because the number of training images is small and the dataset was unbalanced with respect to variabilities in panels, a data augmentation stage was applied to balance this dataset and to increase the sample size. For such purpose, some geometric operations were applied to images; in particular, rotations from $-5°$ to $5°$, and different zooms on the images from -10% to 10%. This augmentation produced a dataset of approximately 5,900 training and validation images. All training, validation and test images were manually labeled (i.e., by marking four rectangle points per panel) using the VGG Image Annotator Tool [3] in order to produce the ground-truth regions were panels were located in the images. Next, the annotated information in each image was stored and adapted to TensorFlow API.

3.2 Global Results

Next, we show some qualitative and quantitative results regarding detection of panels in test images. Figure 3 illustrates several panel detections produced by our method on the same sample images shown in Fig. 1. Note that panels are accurately detected under different variable conditions like: night illumination, perspective, partial occlusions and reduced size in images. Moreover, our approach also worked well in presence of various combined variabilities in images (e.g., small and rotated urban panels).

Fig. 3. Detections achieved by proposed method on images of Fig. 1. Red and green rectangles respectively correspond to ground-truth and detected panels. (Color figure online)

Figure 4 illustrates a correct detection situation when more than one panels are present in the scene.

Fig. 4. Detection results of two panels in a sample image.

Regarding quantitative results, we used the Intersection over Union (IoU) metric [14] as accuracy measure in detections. With this metric, the accuracy of a panel detection is computed by dividing the intersection of respective hand-labeled ground-truth and model-predicted bounding boxes by the corresponding union of these two bounding boxes. The respective computed IoU values for the panels detected in images of Fig. 3 were: 0.93, 0.94, 0.88 and 0.93.

The output of the SSD network returns for each point of an input image the confidence to find a panel centered on that point. In the context of our model we are interested in the absence of False Positives (FP). We think it is preferable, in applications related to advertisements, to leave an old content inside the panel than to crush a wrong part of the image with an updated advertising publicity. Therefore, using the learning sample, we have set a threshold value of 0.7 for such a confidence in a correct panel detection. Thus, if the neuron in the output layer has a value greater than 0.7, we consider that the network decides that there is a panel at that point.

Using this threshold value, an IoU value greater than zero is obtained for all panels in the test images except in one case. Such a case could be considered a False Positive. Table 1 shows the number of posters detected with this threshold, and the IoU rate obtained in each case. This threshold leaves a False Rejection rate of 25% (35 of 140 panels). As can be seen in Table 2, many not detected panels correspond to very small ones (i.e., even much more smaller than 10% of the image size).

On the other hand, it must be determined which IoU threshold can be considered as a correct detection. Figure 5 shows the detection rate that would be obtained (on the y-axis) by setting detections that meet a minimum IoU (on the

x-axis) as correct. It can be seen that with IoU values below 40% the detection rate is 99%. Increasing the IoU threshold beyond 40% reduces the detection success. An interesting value is to consider the detection as correct when the IoU is 80%, in this case our system would also have an 80% success.

Table 1. Number of panels detected and their associated IoU values

IoU	Detected billboards
0.5	2
0.6	1
0.7	2
0.8	10
0.9	41
1	49

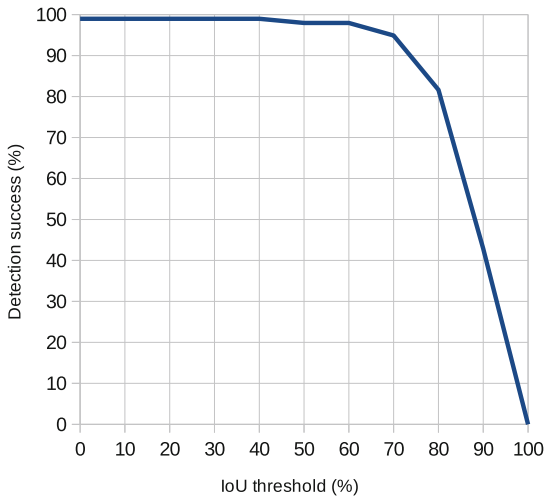

Fig. 5. Percentage of success in detections related to the IoU threshold value.

3.3 Influence of Panel Sizes in Detection Results

Figure 6 shows the distribution ratio of panel sizes with respect to image sizes (in percentages) for the test images. Note that most of panels in images are small (i.e., a 40% of them) and have a size ratio below or equal to 10%; a 81.5% of panels present a ratio size below or equal to 30%; and all panels are smaller or equal than half of the image. In our study, we have not considered very big panels (i.e., above 50% of image size) since their detection becomes easier even with a complex image background.

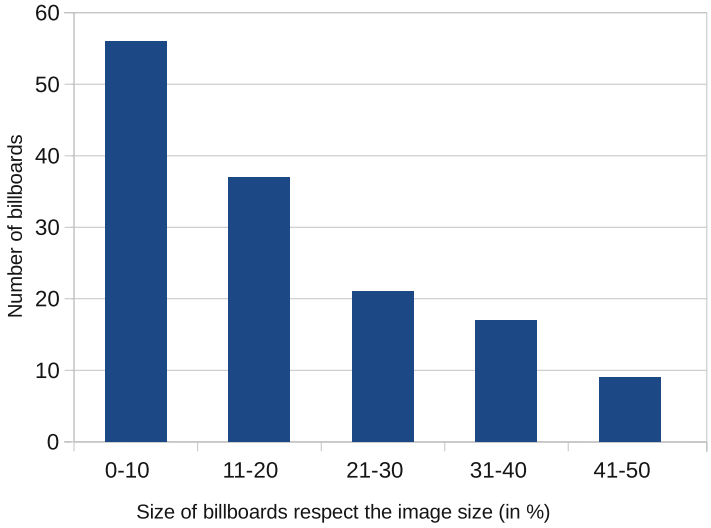

Fig. 6. Number of billboards grouped by different sizes.

With respect to how panel size variability influences its detection, we show in Table 2 that the detection percentage increases as panel area in image increases too. This value grows from around 53.5% in panels that covers less than 10% of the image to 100% when the panel size is in between 41% and 50% of the image. An interesting fact is that when panels are detected in images, their corresponding IoU accuracy is, in most cases, above 85%, regardless the panel size.

Table 2. Achieved detection results related to panel sizes in images.

Size ratio: panel vs image (in %)	Number of panels	Number of detections	Detections percentage	Mean IoU (in %)
0–10	56	30	53.57	86.80
11–20	37	31	83.78	88.54
21–30	21	20	95.24	84.14
31–40	17	15	88.24	87.79
41–50	9	9	100.00	89.74

4 Conclusion

This paper presents a method based on Single Shot MultiBox Detector (SSD) network to accurately locate the position of advertising panels in outdoor urban

images. Our proposal is robust and produces good detections under multiple variabilities like panel sizes, illumination conditions, image perspective, partial occlusion of panels, complex background and multiple panels in scenes. The presented method reduces significantly the number of false detections in test images, which makes it useful in the smart city context. As future work, we aim to improve the detection of very small panels (i.e., those ones producing most false negatives). Another interesting future work consists in recognizing the elements contained inside the panels to determine the brand name of a given advertising or to use the panel detection to update the publicity in an Augmented Reality application.

Acknowledgments. The authors gratefully acknowledge the financial support of the CYTED Network "Ibero-American Thematic Network on ICT Applications for Smart Cities" (Ref: 518RT0559) and the Spanish MICINN RTI Project (Ref: RTI2018-098019-B-100). The third author acknowledge the support of the ESPOL project PRAIM (FIEC-09-2015), the Spanish MICINN Project TIN2017-89723-P and "CERCA Programme/Generalitat de Catalunya".

References

1. Anthopoulos, L.: Understanding Smart Cities: A Tool for Smart Government or an Industrial Trick? Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-57015-0
2. Borisova, O., Martynova, A.: Comparing the effectiveness of outdoor advertising with internet advertising. Bachelor's thesis, JAMK University of Applied Sciences, Finland (2017)
3. Dutta, A., Gupta, A., Zissermann, A.: VGG Image Annotator (VIA) - Version: 1.0.6 (2016). http://www.robots.ox.ac.uk/~vgg/software/via. Accessed 19 Feb 2005
4. Garcia, M., Sotelo, M., Martin, E.: Traffic sign detection in static images using matlab. In: IEEE Conference on Emerging Technologies and Factory Automation (ETFA 2003) (2003)
5. Howard, A.G., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861 (2017)
6. Huang, Y., Hao, Q., Yu, H.: Virtual ads insertion in street building views for augmented reality. In: 18th IEEE International Conference on Image Processing (ICIP 2011), pp. 1117–1120 (2011)
7. Intasuwan, T., Kaewthong, J., Vittayakorn, S.: Text and object detection on billboards. In: 10th International Conference on Information Technology and Electrical Engineering (ICITEE 2018), pp. 6–11 (2018)
8. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
9. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
10. Murty, R., et al.: Citysense: an urban-scale wireless sensor network and testbed. In: 2008 IEEE International Conference on Technologies for Homeland Security (2008)

11. Panchal, T., Patel, H., Panchal, A.: License plate detection using harris corner and character segmentation by integrated approach from an image. Procedia Comput. Sci. **79**, 419–425 (2016). https://doi.org/10.1016/j.procs.2016.03.054
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
14. Rosebrock, A.: Intersection over Union (IoU) for object detection (2016). https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/. Accessed 19 Feb 2005
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Watve, A., Sural, S.: Soccer video processing for the detection of advertisement billboards. Pattern Recogn. Lett. **29**(7), 994–1006 (2008). https://doi.org/10.1016/j.patrec.2008.01.022
17. Wong, D., Deguchi, D., Ide, I., Murase, H.: Vision-based vehicle localization using a visual street map with embedded SURF scale. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8925, pp. 167–179. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16178-5_11