Exploiting Egocentric Vision on Shopping Cart for Out-Of-Stock Detection in Retail Environments

Dario Allegra
allegra@dmi.unict.it

Mattia Litrico
mattia.liltrico@studium.unict.it

Maria Ausilia Napoli Spatafora
m.napolispatafora@studium.unict.it

Filippo Stanco
fstanco@dmi.unict.it

Giovanni Maria Farinella
gfarinella@dmi.unict.it

Department of Mathematics and Computer Science
University of Catania, Italy

Abstract

Continuous detection and efficient monitoring of Out-Of-Stock (OOS) of products in retail environments is a key factor to improve stores profits. Traditional methods require labour-intensive human work dedicated to checking for products to refill raising the requirement of automatic solutions to detect OOS. In this work, we focus on the problem of OOS detection from an egocentric perspective proposing a new weak annotation of the EgoCart dataset. We benchmark the considered challenge employing a deep learning approach for the detection of OOS areas. Specifically, we train a Convolutional Neural Network (CNN) to predict attention maps useful to find OOS in retail areas and hence suggest the retail employers where to intervene. We evaluate results with both objective measures and a subjective analysis provided by human which has reviewed the obtained OOS attention maps. The achieved performance demonstrates that the proposed pipeline is promising to help the refilling process in the retail domain.

1. Introduction

One of the goals of retail stores, such as supermarkets and convenience stores, is to avoid the loss of sales opportunities in order to maximise profits. In this context, monitoring on-shelf availability is a key factor for improving stores profits, because Out-Of-Stock (OOS) of products are often the main reasons for losses. In this paper, the term OOS is used to refer to a visible lack of one or more products on a shelf (i.e, a visible hole in a shelf). To keep high on-shelf availability, store clerks have to continuously walk around stores to monitor and stock up missing products. This activity requires a high human effort, which makes urgent the development of detection methods to automatically report on OOS. Hence, in this paper, we propose to use egocentric video acquired with shopping carts by retail customers with the aim to face the OOS detection problem.

With the availability of big dataset and machine learning solutions, many successful algorithms for object detection have been developed [16][8][17][9], but they usually require fine-grained annotation of a huge amount of data for proper training in a new domain. Since the annotation task is very time-expensive, in the last years weak supervised learning algorithms and weak data annotation have become very attractive [26] [24] [25].

This drove us to propose a new weak annotation for the EgoCart dataset [23] (which includes video frames acquired in a retail environment) to study the problem of OOS of products. We also propose to benchmark the OOS problem by training a Convolutional Neural Network to predict attention maps that highlight OOS in the shelves. Results demonstrate that OOS detection can be automatised using data from an egocentric perspective with good accuracy to help better managing of retails spaces.

The remainder of the paper is organised as follow: in Section 2 we present related work on the OOS detection problem; Section 3 describes the proposed weak annotation for the EgoCart dataset; Section 4 presents the proposed methods; Section 5 reports experimental settings and results; Section 6 reports conclusions and suggestions for future works.

2. Related Work

Scene understanding is a fundamental ability of humans which allow them to explore the world, learn and

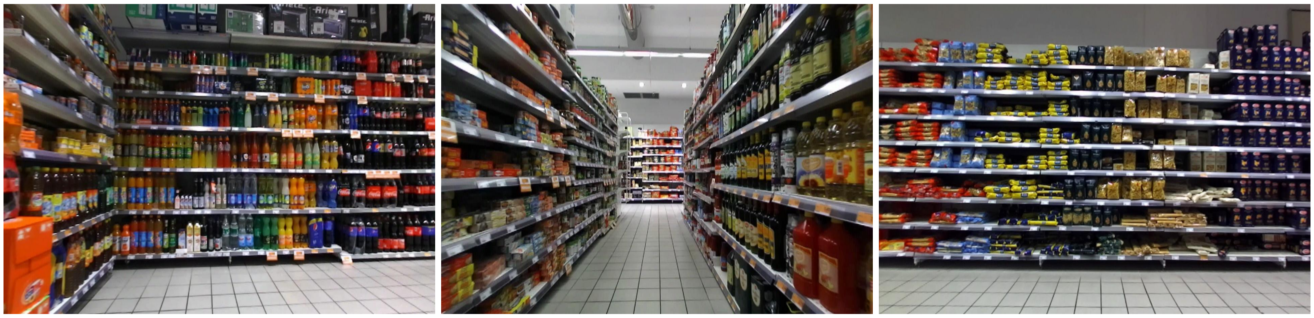

Figure 1. Examples of images belonging to the EgoCart Dataset.

act. Studies have been performed to recognize scene context categories [1], distinguish food vs non-food images [15, 4], detect objects by exploring the domain adaptation paradigm [14], as well as to understand humans from images acquired with an egocentric perspective to recognise personal contexts of interest [6] and to localize them in a store [22].

In retail environments, the main goal is to understand the scene from a customers' point of view in order to infer the behaviour of users during their shopping, and hence better manage the spaces. Recent studies in this context have considered image sequences acquired with cameras mounted on shopping carts carried by retail's customers [20].

In a retail scenario, the continuous monitoring and detection Out-Of-Stock (OOS) of products is a key factor to better manage the spaces and to improve stores profits. In the last decade, several strategies for automatic OOS detection in retail environments have been proposed. Some approaches were quite simple and employed physical sensors or analytical observations [7][12][13]. More recently, image processing techniques have been employed [11][19] to achieve better performance. On the other hand, the most recent works in the field proved that deep learning strategies outperform previous approaches on solving this task [10][3]. However, publicly benchmark datasets to study the OOS problem are still missing.

In 2006, Hausruckinger [7] proposed to count items' sales on a daily basis. The main assumption of the work was that if for a given product no items were sold, then it is to be considered as OOS. In 2007, Ngai et al. [12] proposed to use a Radio-Frequency Identification technology (RFID) to track products in real time. This allowed to monitor the availability of on-shelf products and hence refill OOS. Papakiriakopoulos et al. [13] proposed a machine learning-based approach which used data available in the store information system to check the availability of products.

Moorthy et al. [11] suggested an algorithm based on feature matching which employs Speeded Up Robust Features (SURF) in order to identify and count front-facing products and to reveal OOS. The authors of [19] proposed a supervised learning approach for OOS detection exploiting texture, colour and geometry features together with cascade classifiers and a Support-Vector Machine (SVM) for classification of the potential OOS areas. In 2019, Higa et al. [10] presented a work where shelves were monitored through a CNN to detect and classify the changes in products amount using images acquired by surveillance cameras.

In [3] the authors proposed an Out-Of-Stock detection system which consists of two phases: in the first phase, the objects in the shelves were identified by using Faster R-CNN[17] detector; in the second phase, authors employed classical descriptors based on Canny edge detector, grey level co-occurrence matrix and colour features to perform the actual OOS detection.

Differently from previous works, we propose to deal with the OOS detection problem starting from images captured by a camera mounted in shopping carts. Hence, in this study we employ a new weak annotation of the EgoCart dataset [23] and we benchmark it employing a deep learning based approach to detect OOS areas.

3. Dataset

In this work, we build on the publicly available EgoCart dataset which contains images of a retail environment [23]. Images have been obtained by mounting a camera connected to a laptop on a shopping cart and then extracting frames from the collected videos at a frame rate of 3 fps. Some examples of images belonging to the EgoCart dataset are shown in Figure 1. The dataset contains a total of 19531 images with a resolution of 1280×720 pixels and the related depth maps.

In order to consider the problem of OOS in the retail domain, we have manually re-annotated the EgoCart dataset. We used a weak annotation procedure for labelling each OOS with an approximate position of its centre (see Figure 2). An amount of 8255 RGB images has been annotated with a total of 19079 OOS areas. The remaining 11276 frames of the EgoCart dataset do not contain any OOS and are useful to both provide training signal for those case do not present OOS, as well as for testing the accuracy in terms of false positive for OOS detection. For each frame, we

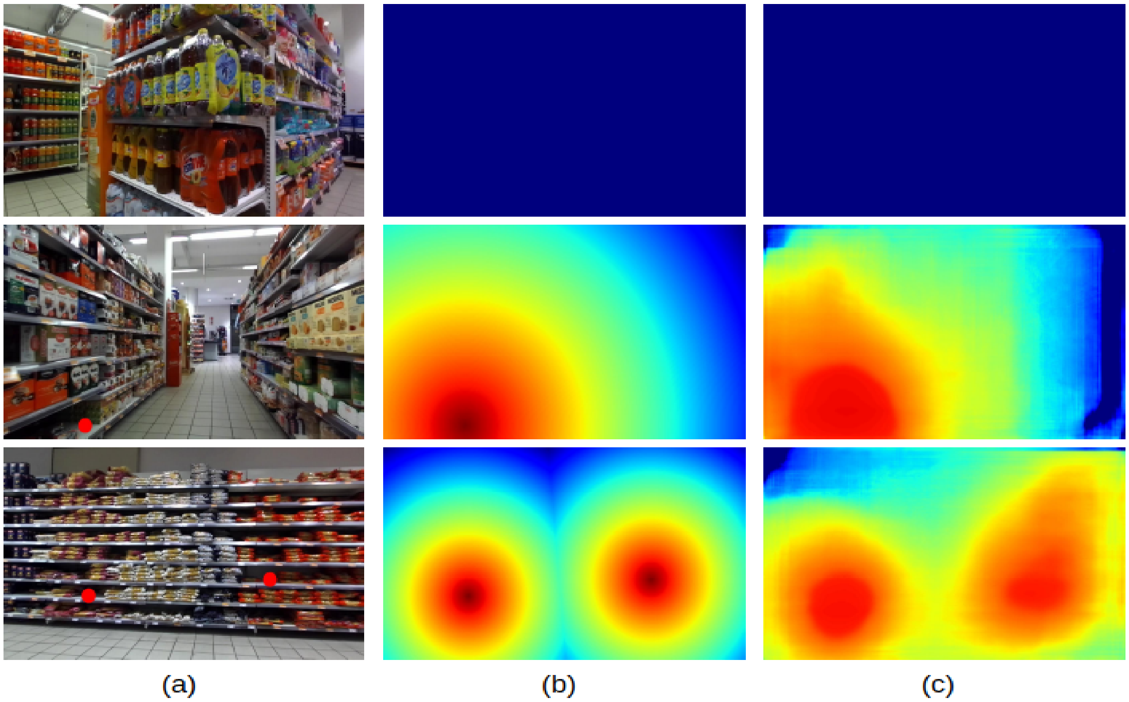

Figure 2. (a) Images annotated with OOS areas (see the red points); (b) Ground truth attention maps obtained using distance transform; (c) Predicted attention maps using the proposed approach

.

have asked two annotators to click approximately at the centre of OOS. Hence, the final annotated dataset presents a set of images annotated with the coordinates of the holes representing an OOS in these images.

It is important to note that since the acquisition have been done in a real retail store scenario, the human manual annotation of OOS results in a very challenging task; this is due to perspective issues, shelf type, lighting and product placement. Thus, the problem of detecting OOS from images reveals to be highly subjective. The proposed OOS dataset is publicly available at the following URL: `https://iplab.dmi.unict.it/OOS/`.

To the best of our knowledge, there is no other annotated dataset to study the problem of OOS detection in retail environment [21].

4. Method

In this section, we describe the approach used to benchmark the task of OOS detection in retail environments. We have explored a Convolutional Neural Network (CNN) based on the U-Net architecture [18]. The CNN was trained to predict attention maps that represent the presence of OOS in certain areas of the shelves. We replace the last layer of the original U-Net with three convolutional layers to obtain single-channel attention maps as output of the network. As ground truth to train the network, we do not use the raw coordinates of the OOS, but a mid-level representation obtained with the followings steps. Firstly, we generate a binary mask where 1 values identify the OOS positions. Then, we compute the distance transform on such binary mask to get a ground truth attention map to be used as a training signal for the CNN.

The distance transform is an image representation based on a distance metric. It converts a binary digital image, consisting of feature (1 values) and non-feature pixels (0 values), into an image where all the pixels have a value corresponding to the distance to the nearest feature pixel [2]. The new intensity value $\mathfrak{D}(\mathbf{x})$ of the element located at position $\mathbf{x}$ of the image is obtained as following [5]:

$$\mathfrak{D}(\mathbf{x}) = \min_{\mathbf{f} \in F} d(\mathbf{x}, \mathbf{f}) \tag{1}$$

where F is the set of feature pixels coordinates and d is a distance function. For our purpose, we use the Euclidean distance as metric d and define F as the set of OOS coordinates.

Then, the generated attention map $\mathfrak{D}$ is normalised in the range [0, 1] as following (see Figure 2):

$$\tilde{\mathfrak{D}}(\mathbf{x}) = \frac{\mathfrak{D}(\mathbf{x}) - min(\mathfrak{D})}{max(\mathfrak{D}) - min(\mathfrak{D})} \tag{2}$$

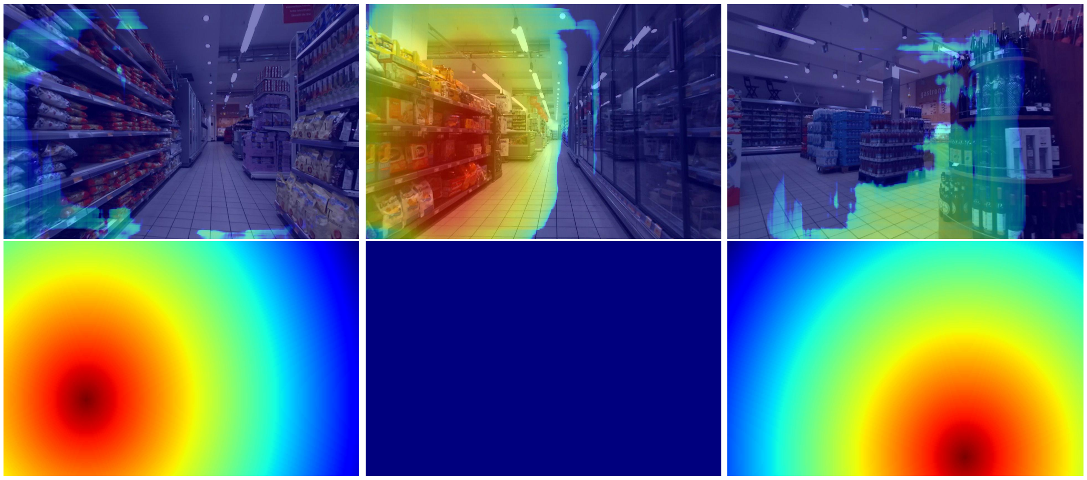

Figure 3. Examples of incorrect predictions (top row) and respective ground truth maps (bottom row).

Pixels colour gradients are proportional to maps intensity values. Examples of attention maps generated in this way from EgoCart images are shown in Figure 2. The top row shows an example where OOS are not present.

5. Experimental settings and results

Images have been rescaled to a size of 320 × 180 pixels and normalised to the range [0,1]. Then, we randomly selected 55% of the images for training, 20% for validation and the remaining 25% for test. We have also employed data augmentation through flipping, colour jitter and cropping patches of 180×140 pixels. For the purpose of training the CNN, we adopted a Mean Absolute Error between predicted attention maps and training maps as loss function. The following hyper-parameters have been used: learning rate 0.001; momentum 0.9; Stochastic Gradient Descent (SGD) as optimizer.

In order to properly evaluate the meaningfulness of the predicted attention maps, we performed both objective and subjective evaluation.

As an objective metric, we used the Root Mean Square Error (RMSE) between the ground truth attention map and the predicted one. The proposed approach achieves an RMSE of 0.139. We propose this RMSE value as a benchmark for future works. RMSE produces a pixel-wise valuation that does not fully capture the attention maps semantic. Hence a subjective evaluation of the results has been performed.

As a subjective evaluation, we define binary voting carried out with the help of human inspection on the test set results. Frames and the respective predictions were analysed side by side in order to check OOS areas with the proposed method and to verify if their existence is correctly suggested in the predicted attention maps. Each of the detected OOS areas has been manually classified as:

Metrics	**Accuracy**	**Precision**	**Recall**
Human Evaluation	78.84%	70.48%	87.99%

Table 1. Results achieved from the Human Evaluation.

	Positive	**Negative**
Positive	70.48%	29.51%
Negative	11.30%	88.69%

Table 2. Confusion matrix obtained from the Human Evaluation.

- *True Positive*: the prediction correctly highlights the existence of the OOS.
- *False Positive*: the predicted attention map suggests an attention region that is not in a true OOS area.
- *False Negative*: the OOS is not detected by the proposed approach.
- *True Negative*: a frame and the related predicted attention map do not exhibit any OOS area.

Table 5 and Table 5 show the results and the confusion matrix of the proposed human evaluation of the results obtained with the proposed deep learning approach.

The lack of similar approaches for OOS detection in literature makes hard a comparison with other state-of-art methods. However, experimental results prove that the proposed approach is promising since it achieves an Accuracy of $78,84\%$ and a Recall of $87,99\%$ which remark that the model is able to detect OOS areas with good performances. This is also confirmed by qualitative results, as shown in some example in Figure 2. For a better qualitative evaluation of the obtained results, the reader can see a demo video at the following link: https://iplab.dmi.unict.it/OOS/. On the other hand, a Precision of $70,48\%$ shows that the CNN is encouraged to produce false positive. Predictions are strongly affected by the high variability in size, lighting and different background of OOS areas. For instance, the perspective distortion at the top shelves is very emphasised and it makes the detection of OOS difficult. Some incorrect predictions are shown in Figure 3. In such cases, the architecture has inaccurate predictions due to the aforementioned issues that make the detection of OOS hard. Moreover, the approach is strongly conditioned by ground truth annotations which in this case is very weak.

6. Conclusions

In this work, we explore the problem of OOS detection from an egocentric perspective in a retail environment by extending the annotation of the EgoCart dataset and proposing to predict attention maps useful to highlight the presence of OOS. To this aim, we used a deep learning architecture based on U-Net. This approach is evaluated objectively and subjectively achieving good performance on the introduced dataset. Future works can be devoted to extend this study by employing different architectures as well as using depth information in order to achieve better performances.

7. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

References

[1] S. Battiato, G.M. Farinella, G. Gallo, and D. Ravi. Scene categorization using bag of textons on spatial hierarchy. In *International Conference on Image Processing, ICIP*, pages 2536–2539, 2008.

[2] Gunilla Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344 – 371, 1986.

[3] J. Chen, S. Wang, and H. Lin. Out-of-stock detection based on deep learning. In *Intelligent Computing Theories and Application*, pages 228–237, 07 2019.

[4] G.M. Farinella, D. Allegra, F. Stanco, and S. Battiato. On the exploitation of one class classification to distinguish food vs non-food images. In *Lecture Notes in Computer Science*, volume 9281, pages 375–383, 2015.

[5] P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 2004.

[6] A. Furnari, G.M. Farinella, and S. Battiato. Recognizing personal contexts from egocentric images. In *IEEE International Conference on Computer Vision*, 2015.

[7] G. Hausruckinger. Approaches to measuring on-shelf availability at the point of sale. Technical report, ECR Europe, 2006.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 03 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 06 2016.

[10] Higa, K. and Iwamoto, K. Robust Shelf Monitoring Using Supervised Learning for Improving On-Shelf Availability in Retail Stores. *Sensors*, 19, 2019.

[11] Moorthy, R and Behera, S. and Verma, S. and Bhargave, S. and Ramanathan, P. Applying Image Processing for Detecting On-Shelf Availability and Product Positioning in Retail Stores. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pages 451–457, 08 2015.

[12] Ngai, E. and Cheng, T. C. E. and Au, S. and Lai, K. Mobile commerce integrated with RFID technology in a container depot. *Decision Support Systems*, 43, 2007.

[13] Papakiriakopoulos, D. and Pramatari, K. and Doukidis, G. A decision support system for detecting products missing from the shelf based on heuristic rules. *Decision Support Systems*, 46:685–694, 02 2009.

[14] Giovanni Pasqualino, Antonino Furnari, Giovanni Signorello, and Giovanni Maria Farinella. An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites. *Image and Vision Computing*, page 104098, 2021.

[15] F. Ragusa, V. Tomaselli, A. Furnari, S. Battiato, and G.M. Farinella. Food vs non-food classification. In *International Workshop on Multimedia Assisted Dietary Management*, pages 77–81, 2016.

[16] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 04 2018.

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 06 2015.

[18] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 10 2015.

[19] Rosado, L. and Gonçalves, J. and Costa, J. and Ribeiro, D. and Soares, F. Supervised learning for Out-of-Stock detection in panoramas of retail shelves. In *IEEE International Conference on Imaging Systems and Techniques*, pages 406–411, 10 2016.

[20] Vito Santarcangelo, Giovanni Maria Farinella, Antonino Furnari, and Sebastiano Battiato. Market basket analysis from egocentric videos. *Pattern Recognition Letters*, 112:83 – 90, 2018.

[21] B. Santra and D. Mukherjee. A comprehensive survey on computer vision based approaches for automatic identification of products in retail store. *Image and Vision Computing*, 03 2019.

[22] E. Spera, A. Furnari, S. Battiato, and G.M. Farinella. Egocentric shopping cart localization. In *IEEE International Conference on Pattern Recognition*, 2018.

[23] E. Spera, A. Furnari, S. Battiato, and G. M. Farinella. EgoCart: a benchmark dataset for large-scale indoor image-based localization in retail stores. *IEEE Trans. on Circuits and Systems for Video Technology*, 2019.

[24] S. Tian, S. Lu, and C. Li. Wetext: Scene text detection under weak supervision. In *IEEE International Conference on Computer Vision*, pages 1492–1500, 10 2017.

[25] Z. Zhao, L. Yang, H. Zheng, I. Guldner, S. Zhang, and D. Chen. Deep learning based instance segmentation in 3d biomedical images using weak annotation. In *Medical Image Computing and Computer Assisted Intervention*, 06 2018.

[26] Z. H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5, 08 2017.