Automatic Check-Out via Prototype-based Classifier Learning from Single-Product Exemplars (Supplementary Materials)

Hao Chen[1,2], Xiu-Shen Wei[1,2,3⋆], Faen Zhang[4], Yang Shen[1], Hui Xu[4], and Liang Xiao[1⋆]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] State Key Laboratory of Integrated Services Networks, Xidian University, China
[3] State Key Laboratory for Novel Software Technology, Nanjing University, China
[4] Qingdao AInnovation Technology Group Co., Ltd, China
{chenh,weixs,shenyang_98}@njust.edu.cn,{zhangfaen,xuhui}@ainnovation.com, xiaoliang@mail.njust.edu.cn

In the supplementary materials, we provide the evaluation metrics, some other qualitative results and the analyses of working mechanism.

1 Evaluation Metrics

The four evaluation metrics used in experiments are introduced in details as follows.

- Check-out Accuracy (cAcc $\in [0, 1]$) is the accuracy when the complete shopping list is predicted correctly, which is computed as

$$cAcc = \frac{\sum_1^N \delta(\sum_{k=1}^K CD_{i,k} = 0)}{N},$$

  where $CD_{i,k} = |P_{i,k} - GT_{i,k}|$ is the counting error for a specific category in an image, $P_{i,k}$ and $GT_{i,k}$ correspond to the predicted and ground-truth number of items in the k-th category in the i-th image, respectively. $cAcc = 1$ means that all items are predicted accurately, *i.e.*, $\sum_{k=1}^K CD_{i,k} = 0$. This is the primary metric.
- Average Counting Distance (ACD) is the average number of counting errors for each image:

$$ACD = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} CD_{i,k}\,.$$

- Mean Category Counting Distance (mCCD) calculates the average ratio of counting errors for each category:

$$mCCD = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i=1}^N CD_{i,k}}{\sum_{i=1}^N GT_{i,k}}\,.$$

⋆ Corresponding author.

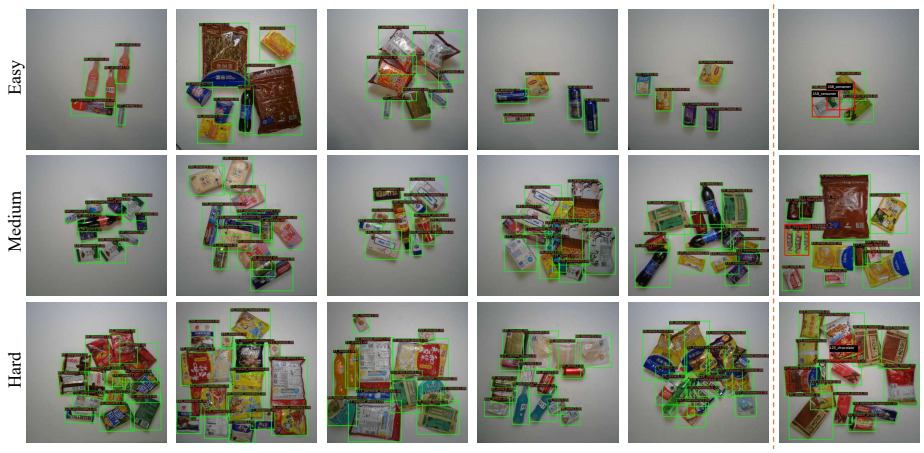

Figure 1. Examples of ACO predictions. Each row denotes visualization results of the check-out image for different clutter modes. The former five columns are the results of the correct predictions, and the latter column corresponds the failure cases. In each figure, the green bounding boxes are the corrected product predictions, while the red bounding boxes are the predictions with wrong labels.

– Mean Category Intersection of Union (mCIoU $\in [0, 1]$) is the overlap between the predicted and ground-truth shopping list, which is defined as

$$mCIoU = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \min(GT_{i,k}, P_{i,k})}{\sum_{i=1}^{N} \max(GT_{i,k}, P_{i,k})} .$$

2 Qualitative Results

In the following, we provide more qualitative results for our PSP method. In Fig. 1, we show the check-out predictions as qualitative results. As seen from the successful results, our PSP method is effective in detecting and locating all categories of products and counting them correctly even at the hard mode. However, there are still some failure cases, which are caused by dense placement and occlusion by other products.

In Fig. 2, we also present the qualitative results of ranking orders before or after soft re-ranking. The first two groups of examples show that the original ranking order is classified incorrectly, and the classification is corrected after our discriminative re-ranking. The third group shows that when the original ranking classification is correct, our re-ranking can retain some fine-grained classes before the background class to ensure the rationality of the ranking. Moreover, the last group shows that when the background category is correctly classified, and the ranking order remains unchanged without re-ranking. Therefore, the last group has no marker box.

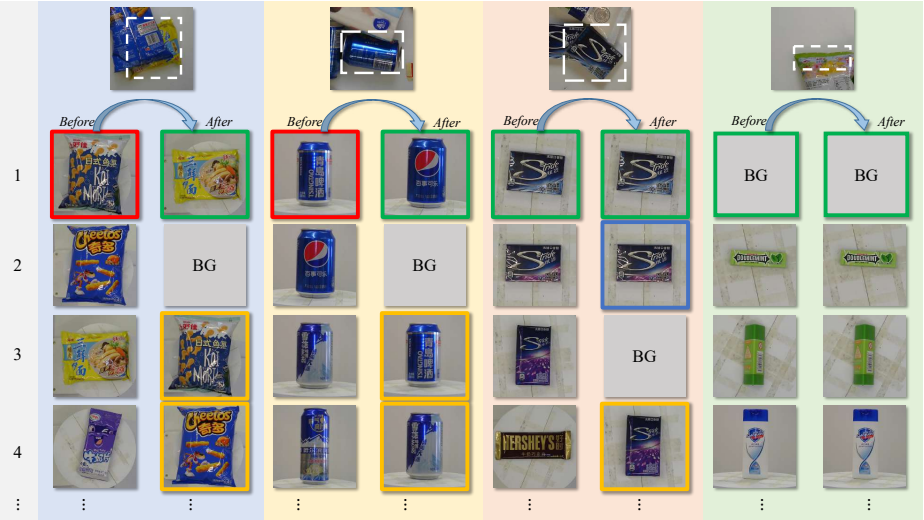

Figure 2. More groups of examples of before and after our soft discriminative re-ranking. In each group, the white dotted box represents the product proposal. The figures with red and green border are wrong and correct predictions, respectively. The figures with blue and yellow border respectively represent that the ranking orders have not and have changed after soft re-ranking. The last group will not be re-ranked due to the correct classification result of the background category, so there is no marker box.

3 Analyses of Working Mechanism

We hereby analyze the working mechanism of our PSP from both qualitative and quantitative perspectives.

For qualitative results, the reason for the well adaptability of PSP to feature diversity can be demonstrated from results in Fig. 3. It shows an illustration of similarity maps between single-product exemplars and product proposals from check-out images. Specifically, we obtain the output $\boldsymbol{F}^{k,i}$ of the i-th exemplar of the k-th category in the last convolutional layer of ResNet-50, as well as the detected product proposal $\boldsymbol{m}_n$ (cf. Eq. (5) of the paper) with a well-trained PSP model, respectively. $\boldsymbol{F}^{k,i} \in \mathbb{R}^{H\times W\times C}$ and a cell of $\boldsymbol{F}^{k,i}$ can be denoted as $\boldsymbol{f}^{k,i}_{h,w} \in \mathbb{R}^{1\times 1\times C}$, where H, W and C are the height, width and number of channels of $\boldsymbol{F}^{k,i}$, h and w are the the vertical and horizontal coordinates of $\boldsymbol{f}^{k,i}_{h,w}$. Then, we can derive the similarity maps between $\{\boldsymbol{f}^{k,i}_{h,w}\}$ and $\boldsymbol{m}_n$ using the corresponding cosine distances. In these similarity maps, the warmer the colors, the smaller the cosine distance, *i.e.*, the higher the similarity. It is obvious that the most similar product exemplar contributes the most to the final ACO classification upon check-out images. Meanwhile, it reveals that our method can adaptively match the various viewpoint product proposals with training product exemplars.

For the quantitative results, we validate the domain adaptation of the prototype-based classifiers without depending on the object detection back-

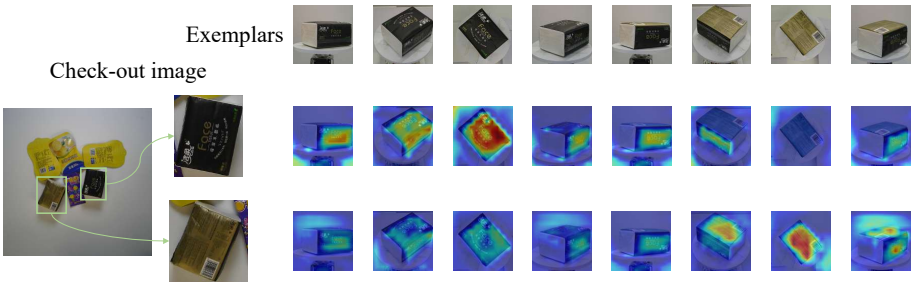

Figure 3. An illustration of similarity maps between single-product exemplars and product proposals from check-out images.

Table 1. The recognition accuracy of two kinds of prototype-based classifiers.

Classifiers	Accuracy	
	Check-out	Exemplar
$\{\mathbf{W}_1\}$ (Exemplar $\rightarrow$ Check-out)	90.89%	54.38%
$\{\mathbf{W}_2\}$ (Exemplar $\rightarrow$ Exemplar)	15.26%	84.08%

bone. In concretely, we use 8 exemplars per category as training data to generate prototypes and learn two kinds of classifiers, *i.e.*, $\{\mathbf{W}_1\}$ derived from exemplars to handle check-out images, and $\{\mathbf{W}_2\}$ from exemplars to still exemplar images. Then, we employ $\{\mathbf{W}_1\}$ and $\{\mathbf{W}_2\}$ to directly perform upon the ground truth product bounding boxes of check-out images (for removing the localization influence) and product exemplars for recognition accuracy comparisons. In another words, $\{\mathbf{W}_1\}$ and $\{\mathbf{W}_2\}$ are both generated from the single-product exemplar prototypes, but their learning objectives are designed to recognize check-out images and exemplar images, respectively. Different learning objectives brings different recognition abilities. As shown in Table 1, even $\{\mathbf{W}_1\}$ learned from exemplars, thanks to our prototype-based classifier generation, it can overcome domain gaps and achieve better results for check-out images (90.89%) than for exemplars (54.38%). Regarding $\{\mathbf{W}_2\}$, since the generation goal is for recognizing exemplars, it is reasonable that $\{\mathbf{W}_2\}$ works worse for check-out images (15.26%) than product exemplars (84.08%).