Toward Retail Product Recognition on Grocery Shelves

Gül Varol[1,3], Rıdvan Salih Kuzu[2,3]

[1]Departments of Computer Engineering and [2]Electronical and Electronics Engineering, Boğaziçi University
34342, Bebek, İstanbul, Turkey

[3]İdea Teknoloji
34398, Maslak, İstanbul, Turkey

{gul.varol, ridvan.salih}@boun.edu.tr

ABSTRACT

This paper addresses the problem of retail product recognition on grocery shelf images. We present a technique for accomplishing this task with a low time complexity. We decompose the problem into detection and recognition. The former is achieved by a generic product detection module which is trained on a specific class of products (e.g. tobacco packages). Cascade object detection framework of Viola and Jones [1] is used for this purpose. We further make use of Support Vector Machines (SVMs) to recognize the brand inside each detected region. We extract both shape and color information; and apply feature-level fusion from two separate descriptors computed with the bag of words approach. Furthermore, we introduce a dataset (available on request) that we have collected for similar research purposes. Results are presented on this dataset of more than 5,000 images consisting of 10 tobacco brands. We show that satisfactory detection and classification can be achieved on devices with cheap computational power. Potential applications of the proposed approach include planogram compliance control, inventory management and assisting visually impaired people during shopping.

Keywords: Object recognition, object detection, retail product, grocery image, bag of words

1. INTRODUCTION

Interpretation of grocery images for extracting meaningful information is a growing topic in computer vision. Manufacturers invest resources to check planogram compliance in groceries. The automation of this process is of great importance for inventory management because it is both time saving and more reliable than manual control. Besides this optimization application, product analysis from images can guide a visually impaired user through a grocery.

In the literature, object detection and recognition are well established problems. However, some problems such as low resolution images, occlusions, scale variances, large sizes of object classes make the recognition difficult. On the other hand, there is the problem of localization which is referred as object detection. We may need to search through a large image to determine where the object is found; however, it may require high complexity depending on the classification algorithm.

A real-time product detection system from video is presented in [2]. Some effort for matching database images on an input image is shown in [3] by using scale-invariant feature transform (SIFT) [4] vectors in an efficient manner. Another study focuses on logo detection in natural scenes by spatial pyramid mining [5]. In a patent [6], authors apply planogram extraction based on image processing by using a combination of several detectors. SIFT matching and optical character recognition are some of them.

In image classification, recent studies show that dense sampling of SIFT descriptors outperform sparse sampling (i.e. sampling from the interest points) [7]. Since dimensionality of SIFT vectors are very high when combined, we need a quantization approach such as k-means clustering. Using bag of visual words technique originated from text retrieval applications, we form a descriptor by counting the occurrences of visual words. Recent studies using such method show success especially in object scene classification [8, 9, 10].

In our system, we follow the bag of visual words approach for brand classification. We combine shape and color information by building separate vocabulary for each aspect as in [11]. We investigate the effect of vocabulary size for both aspects.

In order to decrease the search space, we segment the image into products and non-products. Here, we make use of the similarity between classes and call this phase product detection. For each detected product, we apply brand classification. In our case, we work with tobacco products. They have high similarity between classes in terms of shape and design. The part which distinguishes the brands from each other is usually low resolution when extracted from the whole shelf image.

The rest of this paper is organized as follows. Section 2 presents our dataset and the data collection procedure. Section 3 details our approach and methodology to solve product recognition problem. In Section 4, we report our results on the dataset and finally discuss our findings in Section 5.

2. DATASET

We constructed a database of shelf and product images with several properties. In order to have lightning and shelf design variance, we visited around 40 groceries. To be able to examine the effect of the capturing quality, we took photographs with 4 different cameras: one iPhone5S, one iPhone4, one Nikon Coolpix S3 and one Sony Cyber-shot® DSC-W300. In total, we took 354 shelf images. To further test shelf segmentation performance and to have shelf images taken from various distances, we covered different numbers of shelves on an image. The distribution over the number of shelves is summarized in Table 1. Furthermore, we took multiple photographs of the shelves in a particular rack by sliding the frame one shelf down at each time. Thus, we enable a future study on image stitching such as constructing the image of the whole rack by merging multiple small shelf images. In our annotation, we noted all this information. A sample group of images of the same rack is shown in Figure 1.

Table 1. Distribution of grocery images over the number of shelves

Number of shelves	2	3	4	5	6	7
Number of images	92	137	83	34	6	2

Each shelf image is annotated by drawing bounding boxes around product packages using Image Clipper[1] utility. In total, there are around 13,000 products on 354 images. Each image took about 1.5 minutes to annotate. We selected 10 brand classes and noted the products that belong to those classes. On the average 200 products per class are found on the images. The rest 10,000 products did not belong to any category; therefore may be considered as negative class. The brand classes are presented in Figure 2.

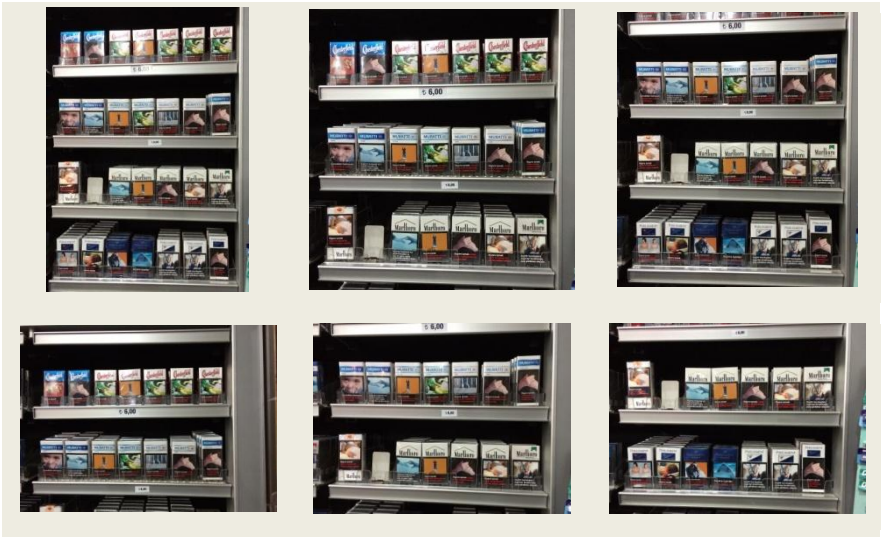

Figure 1. A sample rack of 4 shelves photographed as a whole (top-left) and as in pieces (the rest)

[1] code.google.com/p/imageclipper/

Figure 2. Overview of brand classes

Figure 3. Various images of one brand

We further took photographs of single products on a controlled environment. From each class, 5 product packages are photographed by 4 different cameras multiple times in different groceries on a white background. We paid attention to having both light and dark, noisy and clear, planar and slightly angular versions of a single product image. Sample images of the same product are illustrated in Figure 3. To accelerate the photographing step, we took 4 products' image at a time; therefore, we needed to crop them out. For this purpose, we developed an image processing program to segment the rectangular products from the flat background. Basically, we applied Gaussian blur, Canny edge detection and dilation. On this pre-processed image, we fit first a polygon, then a bounding rectangle and obtain the boundaries of the products. We extracted around 3,600 product images with this technique and eliminated the unsuccessfully segmented images. Manual cropping could be more accurate; however, it would take on the average 5 products per minute, which would make 12 hours of work.

In our experiments, we use controlled product images for training and product images cropped from shelves for testing. This type of dataset splitting is preferred to test whether images taken in a controlled environment are usable for learning the real world images. Furthermore, we construct a subset of the dataset for our experiments due to the unequal distributions over classes. We consider the class with the minimum number of images to determine the number of instances. Finally, we obtain 274 instances per class in training and 100 instances (except three classes with fewer instances) per class in testing.

3. METHODOLOGY

3.1 Segmentation

3.1.1 Product Detection

Instead of searching for a logo everywhere on the shelf image, we extract the structure of the market shelves prior to applying object recognition. At this stage, we apply a top-down approach rather than bottom-up. Here, we make use of the similarity between classes in terms of shape. A special case with tobacco products is that they have the same warning images across brands, but in many cases there are some between-class similarities. For instance, different beverage brands have similar bottle structure. We have the assumption that the products will not have much out-of-plane rotation and will have similar aspect ratio, which is the case on the grocery shelves.

Considering this similarity, we propose training our own custom cascade object detection which is known to yield fast detection. It consists of a cascade of boosted classifiers working with histogram of oriented gradients (HOG) features. This framework is originally developed for face detection problem [1], but it is known to be convenient to detect a variety of object classes. We provide product images regardless of their brands as the positive samples and images cropped from the background as the negative samples.

Our knowledge about the context of the shelf structure enables us to constrain the detection. Many false positives occur after applying detection. To avoid very small and very big detection results, we set our minimum and maximum desired height adaptively according to the computed number of shelves. How we estimate the number of shelves is detailed in the next subsection. Let N be the number of shelves and h denote the height of the shelf image. Then we can set the maximum height threshold to h/N, which means that a product cannot be larger than the shelf height. Similarly,

we set the minimum height threshold to *5h/N*, meaning that a product should cover at least one fifth of the shelf height. This improves our product detection performance significantly by removing false positives.

To treat the false positives that further satisfy the threshold constraints, we check for the outliers. By computing the mean width and height of the detections, we get information about the approximate size of the products because they are of similar size once taken from the same distance. We discard the detections with width or height 2σ away from the mean width or height where σ denotes the standard deviation. This procedure also improves the performance. A typical product detection result is shown in Figure 4.

As a result of product detection module, we obtain the bounding boxes of the products on shelves in a fast manner which is executable as real-time on devices with limited resources such as smart phones. The next step is to recognize which brand logo appears on each product.

Figure 4. Product detection demonstration overlaid with shelf detection

3.1.2 Shelf Detection

In addition to the segmentation of the image into product and non-product, we propose a technique to determine shelf boundaries. This is not a mandatory but useful step. Once we know the shelf boundaries, we can count them and estimate the number of shelves.

We compute the histogram of the products' projection on the y-axis to examine the distribution over the shelves. Figure 5(top) illustrates a sample histogram computation. The peaks make the shelf segmentation obvious for the human eye. In order to get rid of the noise produced by the false positives, we apply Gaussian filter to the signal. The smoothed signal can be seen in Figure 5(bottom). On this signal, we label the peaks automatically as the product locations and their midpoints as the shelf boundaries. Our previous work [12] which relies on Hough transform for line detection is not robust against variations in shelf design and is computationally more complex.

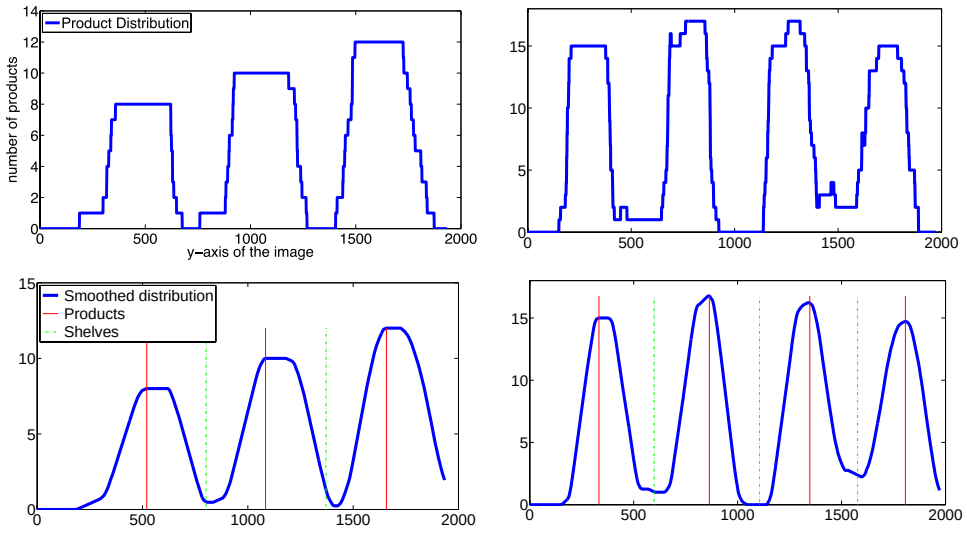

Figure 5. Shelf boundary computation for the two examples in Figure 4

3.2 Brand Recognition

For the brand recognition through logo images falling inside the detected regions, we prefer bag of words technique. The detected region contains both the logo and the warning image. Since the warning part is common across brands, we consider the first 40% portion of the image from the top for classification. Similarly, we use only logo parts of the product images in the training phase. The product detection module does not guarantee a perfect framing for the product. Some part of the logo may not be covered. Moreover, products can be partly occluded by the shelf, the price sticker etc. Therefore, using the frequency of local descriptors is a convenient approach.

We represent the logo image by combining its shape and color information. One aspect itself is enough to discriminate brands up to an extent. When two brands have very similar shape, they may differ in color or vice versa. For this reason, we get better recognition performance with the combination. We use SIFT features for shape description and HSV values for color description.

The conditions from one grocery to another may vary significantly. The relative position of the shelves according to the lamps or the amount of sunlight that enters the grocery affect the appearance of the products on the image. Moreover, if the place is dark, one may need to use camera flash, which entirely captures different colors. One another difficulty occurs because of light reflection when the products are covered with cellophane or when the products are placed in a glass cabinet.

The color space we work is HSV due to its better invariance to illumination. We obtain a color vocabulary by clustering 3-dimensional HSV values gathered from the training images using k-means. Each cluster center in the vocabulary represents a visual word. Given a set of V_c visual words, each image is then represented by a V_c-dimensional normalized frequency histogram. In order to later assign a color value to a visual word, we search for the nearest cluster center using k-d tree structure.

SIFT feature vectors are densely sampled from a regular grid overlaid on the image. These vectors are clustered to construct a shape vocabulary of size V_s in the same way as the color vocabulary.

For each image, we have one histogram for color and one histogram for shape. We concatenate two normalized histograms to obtain a joint frequency histogram. However, depending on the numbers of clusters, V_c and V_s, one modality may dominate the other if it has larger dimensionality. To handle this issue, we perform a weight learning approach. Let w_c and w_s be the weights associated with color and shape descriptor respectively. Then,

$$f(I) = \begin{bmatrix} w_c f_c(I) \\ w_s f_s(I) \end{bmatrix}$$

gives the final feature vector *f(I)* for image *I*. Here, $f_c(I)$ and $f_s(I)$ denote the frequency histograms for color and shape aspects respectively. Since the final vector will be normalized, one free parameter is to be learnt.

Classification is performed with a multi-class SVM with Gaussian radial basis function as the kernel. The hyperparameters of the classifier are optimized on a 10% subset of the training set with 10-fold cross validation.

4. EXPERIMENTS

The experiments are carried out on our dataset presented in Section 2. Both product detection and brand recognition performances are evaluated. Furthermore, we investigate the effect of vocabulary size for both color and shape to choose a convenient *k* in k-means. Other parameters related to SIFT computation are also examined. These include the patch size *P*, and the grid spacing *G*. All brand images are resized to have the same height. Originally, the training images are much larger than the test images since test images are cropped from shelf images.

Product detection module was trained with 375 randomly selected product images taken in controlled environment as the positive samples. In addition to positive samples, we provided 198 negative samples, namely images taken from the background, product boundaries and shelf boundaries. A product is considered as true positive considering the intersection area between the detection and the ground truth. If the inner intersection is above a certain threshold and the outer intersection is below a certain threshold, we accept the detection. We also accept only one detection per product. Therefore, overlapping detections decrease the performance. The system is able to detect the products on 354 shelf images with 94% recall and 81% precision. In total, there exist 13,184 products on these images. The improvements through aforementioned constraints can be seen in Table 2.

Table 2. Performance values for product detection module

	Recall (%)	Precision (%)
no constraint	89	69
min-max height	94	75
width-height deviation	94	81

The high recall rate of the product detection is a promising result. A complete system should perform brand classification on each detected region. However, for now in this study, we do not handle products which do not fall under any category. A system which deals with this problem can automatically ignore the false alarms. Therefore, we are interested more in recall than precision.

We test the performance of brand recognition on already cropped images of the products that are from one of the classes we have selected. Before applying fusion, we evaluated the performances of the individual aspects separately. Using only shape features, 85.9% classification performance is obtained with $P = 30, G = 25$ and $V_s = 400$ on images resized to fixed height of 100 pixels. The effects of grid spacing, patch size and shape vocabulary size can be observed from Figure 6. We see that grid spacing has high impact on accuracy. On the other hand, patch size and vocabulary size do not alter the accuracy significantly.

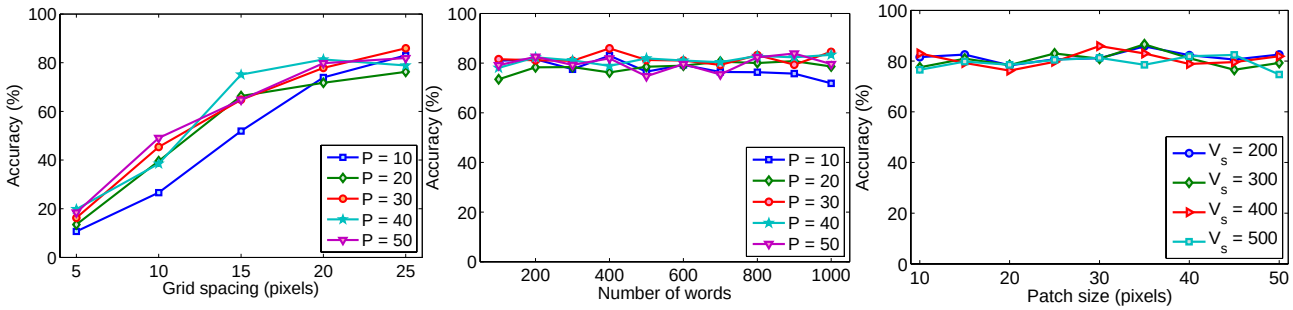

Figure 6. Accuracy results using only SIFT features. Left: The effect of grid spacing for $V_s = 400$ Middle: The effect of vocabulary size for $G = 25$. Right: The effect of patch size for $G = 25$.

Using only color features, 60.5% classification performance is obtained with $V_c = 50$ on images resized to fixed height of $r = 150$ pixels. The effect of the color vocabulary size and the resizing is summarized in Figure 8(a)(b). We do not observe a clear change in accuracy for different values of vocabulary size and resizing. However, we use a much smaller vocabulary for color than shape since the dimensionality of color is only 3.

With the selected parameters from each aspect, we train a combined model to learn a better classification. We experiment both feature-level and score-level fusion for combination. Simple concatenation is applied for feature-level fusion. Averaging the probability outputs of two SVM's is applied for score-level fusion. We search for the best weighting (w_c, w_s) between two aspects and find $w_c = 0.1$ empirically which results in 92.3% accuracy with feature-level fusion. Figure 8(c) shows the results of different weightings. Although score-level fusion results in similar and even better accuracy for some values of w_c, we prefer to use feature-level fusion since scores provided by SVM are calculated using the distances from the margins and they do not necessarily reflect proper probabilities. We notice that shape has a higher weight than color for distinguishing these 10 product categories. As expected, the combined classification results in better accuracy than both individual classifications. The confusion matrix is shown in Figure 7.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_1	96	0	0	1	0	0	0	0	0	3
C_2	0	95	0	1	0	2	0	0	0	2
C_3	0	1	65	0	0	0	0	0	0	1
C_4	1	7	0	83	0	0	3	1	1	4
C_5	0	0	0	0	93	2	1	0	1	3
C_6	1	0	0	0	0	95	1	0	2	1
C_7	0	0	0	0	0	0	100	0	0	0
C_8	0	0	0	0	1	0	0	98	0	1
C_9	0	0	0	0	0	0	0	0	62	13
C_{10}	0	0	0	0	6	3	1	0	6	59

Figure 7. Confusion matrix for brand classification

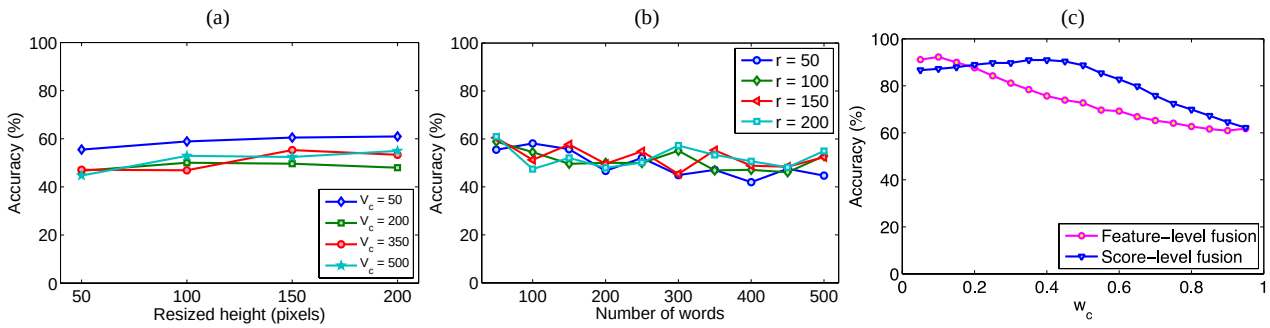

Figure 8. Accuracy results using only HSV features (a)(b) and using both SIFT and HSV (c).

5. CONCLUSIONS AND FUTURE WORK

This work presents an approach for retail product recognition on grocery shelves and introduces a novel dataset which enables researchers to study various aspects of shelf images. We propose to first segment the image to infer product locations and to then classify logos into one of the predefined brands. We quantitatively evaluate the effectiveness of our method on images taken in real world settings. The proposed framework has two separate modules which work individually with satisfactory performances. In a future work, products from undefined categories should be handled in order to have an overall performance measure and a working prototype. Moreover, we have not applied any preprocessing such as noise reduction or perspective correction in this study. Such techniques will also allow us to improve our system in the future.

ACKNOWLEDGMENTS

This research was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under TEYDEB grant 3130322. The authors would like to thank Yusuf Sinan Akgül for his support and assistance.

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," CVPR, 511–518, (2001).

[2] A. Auclair, L. D. Cohen and N. Vincent, "How to use SIFT vectors to analyze an image with database templates," Adaptive Multimedia Retrieval, ser. Lecture Notes in Computer Science, N. Boujemaa, M. Detyniecki and A. Nürnberger, Eds., vol. 4918. Springer, 224–236, (2007).

[3] T. Winlock, E. Christiansen and S. Belongie, "Toward real-time grocery detection for the visually impaired," CVPRW, 49 –56, (2010).

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, 91–110, (2004).

[5] J. Kleban, X. Xie and W. Y. Ma, "Spatial pyramid mining for logo detection in natural scenes," IEEE International Conference on Multimedia and Expo (2008).

[6] A. Opalach, A. Fano, F. Linaker, and R. Groenevelt, "Planogram extraction based on image processing," Patent US 8 189 855, (2012).

[7] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," ICCV (2005).

[8] J. Sivic, B. Russell, A. Efros, A. Zisserman and W. Freeman. "Discovering object categories in image collections," ICCV (2005).

[9] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars and L. Van Gool. "Modelling scenes with local descriptors and latent aspects," ICCV (2005).

[10] G. Dorkó and C. Schmid. "Selection of scale-invariant parts for object class recognition," ICCV (2003).

[11] M.-E. Nilsback and A. Zisserman, "A Visual Vocabulary for Flower Classification," CVPR (2006).

[12] G. Varol, R. S. Kuzu and Y. S. Akgül, "Product placement detection on shelves based on image processing," IEEE Signal Processing and Communications Applications Conference (2014).